EFFICIENT ESTIMATORS FOR EXPECTATIONS IN NONLINEAR PARAMETRIC REGRESSION MODELS WITH RESPONSES MISSING AT RANDOM & DATA INTEGRATION IN HIGH DIMENSION WITH MULTIPLE QUANTILES

A Dissertation

by

GUORONG DAI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ursula Müller
Co-Chair of Committee,	Raymond Carroll
Committee Members,	Jeffrey Hart
	Yonghong An
Head of Department,	Jianhua Huang

August 2019

Major Subject: Statistics

ABSTRACT

This dissertation contains the two research projects in my Ph.D. study.

The first project considers nonlinear regression models that are solely defined by a parametric model for the regression function. The responses are assumed to be missing at random, with the missingness depending on multiple covariates. We propose estimators for expectations of a known function of response and covariates. Our estimator is a nonparametric estimator corrected for the regression function. We show that it is asymptotically efficient in the Hájek and Le Cam sense. Simulations and an example using real data confirm the optimality of our approach.

The second project deals with aggregating and analyzing high dimensional data, which come from multiple experiments and thus have different responses, but share the same predictors. The measurements of the predictors may be different across experiments. In each experiment multiple conditional quantiles are considered simultaneously, assuming a linear relationship between the response and predictors. To select the predictors that affect any of the responses at any of the quantile levels, we propose a penalized estimation process and an information criterion and study the asymptotic properties. Simulations and a real data application demonstrate the advantage of combining information from multiple experiments and quantile levels.

DEDICATION

To the twentieth century

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Dr. Ursula Müller, for her guidance and support throughout my doctoral studies. Dr. Müller has been a source of constant personal encouragement and intellectual stimulation not only during the research work but over the four years I have been at Texas A&M. It is my honor to be able to work with her.

My thanks also extend to the other members of my committee, Dr. Raymond Carroll, Dr. Jeffrey Hart and Dr. Yonghong An, for their interest and insightful comments.

I want to thank everyone in the Department of Statistics, including faculty, staff, visitors and my fellow graduate students. I feel fortunate to spend my four years in such a friendly and stimulating environment. I am also thankful to my friends outside the department, especially Xiaobai, Mincheng, Yingyin and Mengmeng. Life out of work is much more colorful with them.

Finally, I dedicate this work to my family: my parents, Dr. Hong Dai and Ms. Wen Wu, and my father's sister and her husband, Prof. Yan Dai and Prof. Zhaoguang Ge, for their sacrifices and devotion. Their unconditional love and endless support have always been my strength.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised and supported by a dissertation committee consisting of Dr. Ursula Müller, who is my advisor, Dr. Raymond Carroll, who is my co-advisor, Dr. Jeffrey Hart of the Department of Statistics, and Dr. Yonghong An of the Department of Economics.

The data analyzed in Section 3.4 were provided by Dr. Xin Gao of the Department of Mathematics and Statistics at York University, Canada.

All other work conducted for the dissertation was completed by myself independently.

Funding Sources

The second project in the dissertation is based on joint work with Dr. Raymond Carroll, whose research was supported by a grant from the National Cancer Institute (U01-CA057030).

TABLE OF CONTENTS

Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction and literature review of Chapter 2

In Chapter 2 we study efficient estimation of expectations in a nonlinear regression model that is defined solely by the conditional constraint

$$E(Y|X) = r_\vartheta(X), \quad \vartheta \in \Theta \subset \mathbb{R}^p \tag{1.1}$$

and therefore also known as conditional mean model. Here the regression function r_ϑ is assumed to be known up to a parameter vector ϑ and X is a d-dimensional random vector. The nonlinear regression model is an important model for applications; see, for example, the books by Bates & Watts (1998) and Seber & Wild (1989).

In the literature it is quite common to to introduce a third variable $\varepsilon = Y - r_\vartheta(X)$, especially if the covariates X and errors ε can be assumed to be independent; see, for example, Wang & Rao (2001), which studies linear regression with missing responses. We do not make the independence assumption: in many situations, especially in applications in Econometrics, model (1.1), which we consider here, is more suitable because of its flexibility.

We are interested in the scenario when responses Y are possibly missing and work with an indicator variable Z that is 1 if a response Y is observed and 0 if it is missing. Our sample consists of independent copies (X_i, Z_iY_i, Z_i), $i = 1, ..., n$, of a base observation (X, ZY, Z). The indicator Z tells us if a zero response is a numerical zero or a missing value. More specifically, we assume that the responses are missing at random (MAR), i.e., the probability that Y is missing depends only on the covariate vector X that is always observed,

$$\mathrm{pr}\{Z = 1|X, Y\} = \mathrm{pr}\{Z = 1|X\} = \pi(X).$$

The MAR assumption is quite common in applications; see, for example, the book by Little &

Rubin (2019). It in particular implies that Z and Y are conditionally independent given X.

Our goal is to efficiently estimate expectations $E\{h(X,Y)\}$ of the joint distribution in model (1.1), where h is some known square-integrable function. This is a quite general problem: we basically estimate the entire joint distribution of the vector (X,Y). In the literature usually only estimation of the mean response $E(Y)$ is considered; see, for example, Matloff (1981), Cheng (1994), Wang & Rao (2001), Wang & Rao (2002) and, for further references, Müller (2009). Other examples of such expectations are moments of Y or X, mixed moments, and probabilities involving X and Y such as $P(X < Y)$. Estimation of $E\{h(X,Y)\}$ is also considered in Müller (2009) in the more restrictive nonlinear regression model with independent covariates and errors. Müller (2009) exploits the independence assumption by writing $E\{h(X,Y)\}$ as a convolution integral, which can be estimated in a relatively straightforward way. Since the distribution of the errors in our model depends on the covariates, our approach is quite different.

An obvious approach to estimate $E\{h(X,Y)\}$ in the missing data model with MAR responses is to use the Horvitz-Thompson family of estimators

$$n^{-1}\sum_{i=1}^{n}\frac{Z_i}{\widehat{\pi}(X_i)}h(X_i,Y_i),$$

where $\widehat{\pi}(\cdot)$ is an estimator of the probability $\pi(\cdot)$ from above. Hirano et al. (2003) prove its root-n asymptotic normality in a binary treatment model when $\pi(x)$ is estimated by the series logit method. We will consider this estimator in Section 3.3 and compare it with our method.

As in Müller et al. (2006), who discuss estimation of expectations $E\{h(X,Y)\}$ in a simple linear regression model, we use a nonparametric estimator $\widehat{H}_{np}$ and improve it by adding a correction term $\widehat{\Gamma}$ that takes the nonlinear structure into account,

$$\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}, \tag{1.2}$$

with $\widehat{\Gamma}$ defined in equation (2.1) in Section 2.1. Our nonparametric estimator $\widehat{H}_{np}$ for the first part

of (1.2) is a partially imputed estimator,

$$\widehat{H}_{np} = \frac{1}{n}\sum_{i=1}^{n}\{Z_i h(X_i, Y_i) + (1 - Z_i)\widehat{\chi}(X_i)\}, \tag{1.3}$$

where $\widehat{\chi}(x)$ is the Nadaraya-Watson estimator of $\chi(x) = E\{h(X,Y)|X = x\}$, similarly as in Cheng (1994) (see Section 2.1.1 for details). Alternatively one could, as in Cheng (1990), use a full imputation approach, which also replaces observed cases with estimators. In the nonparametric model full imputation and partial imputation are asymptotically equivalent (see Cheng, 1994), which is intuitively clear since the model contains no structural information. For this article we prefer partial imputation, for reasons of speed and simplicity.

We will show that the estimator proposed in this paper is efficient in the sense of Hájek and Le Cam. The efficiency results imply asymptotic normality, which is useful for constructing approximative confidence intervals for expectations $E\{h(X,Y)\}$ of known square-integrable functions $h(X,Y)$. To the best of our knowledge, our estimator is the first efficient estimator for $E\{h(X,Y)\}$ in the parametric MAR multiple regression model (1.1). Müller et al. (2006) proposes an efficient estimator for univariate linear regression, but does not provide technical details. The results of this paper also apply to the usual model with no missing data, i.e., when all indicators equal one and $\pi(\cdot) \equiv 1$, so this is covered as a special case.

Chapter 2 is organized as follows. In Section 2.1 we provide a complete and detailed derivation of the stochastic expansion of the nonparametric estimator and of the correction term. Section 2.2 characterizes efficient estimators of functionals of the joint distribution and gives the efficient influence function for estimating $E\{h(X,Y)\}$ in our model. The efficiency of our estimator is established by showing that the expansion in Section 2.1 matches the efficient influence function in Section 2.2. In Section 2.3 we explain how our estimator can be implemented and compare it with other methods in various scenarios, using computer simulations. The results are positive throughout and confirm the theoretically proved optimality of our approach. In Section 2.4 we illustrate our approach by means of a real data set. Section 4.1 concludes Chapter 2 and discusses

further questions. All the technical details can be found in Appendix A.

1.2 Introduction and literature review of Chapter 3

Chapter 3 considers data integration across multiple responses with high dimensional covariate when the underlying models are based on quantile regression.

To set the stage for this work, first consider K linear regression models

$$Y_k = X_k^{\mathrm{T}} \alpha_k^* + U_k, \quad k = 1, \ldots, K, \tag{1.4}$$

where Y_k is a scalar response, X_k is a p-dimensional deterministic predictor, α_k^* is a p-dimensional parameter vector, and U_k is the error term in each model $k \in \{1, \ldots, K\}$. Zellner (1962) referred to this set of models as *seemingly unrelated regressions* and proposed the idea of estimating the regression parameters simultaneously using a generalized least squares method. We are interested in the situation where these linear regressions have dependent responses and share the same set of predictors. The design matrices may be different across the models. This is, for example, given if individuals are assessed through various responses from different experiments, so that the values of covariates differ among the experiments, while the predictors are the same in all experiments (Gao & Carroll, 2017).

To meet the high-dimensional data situation, we allow the dimension of the parameter vector $p = p_n$ to tend to infinity as the sample size n increases. In addition, we assume that the data are sparse, i.e., most of the parameters are exactly zero, which means that only a fraction of the predictors significantly affect the response.

An important goal is to identify the relevant predictors. One possible approach is to aggregate each predictor's effect in all experiments by forming groups; see Gao & Carroll (2017), who developed a group penalized estimation method using a pseudolikelihood. To handle the unspecified dependence between the responses in the K experiments, they pooled the marginal likelihoods and imposed L_2-group penalization on the grouped parameters. The group penalty was introduced in a 1999 Australian National University Ph.D. thesis by S. Bakin and then applied to group selection

questions by Yuan & Lin (2006). Gao & Carroll (2017) used it to select predictors that are influential across experiments. The main tool in that article is the smoothly clipped absolute deviation penalty, which was proposed in Fan & Li (2001). In addition, the authors used the concept of the Bayesian information criterion to also develop a pseudolikelihood information criterion that applies to the high-dimensional scenario. Although the pseudolikelihood approach is an important advancement, it suffers from the fact that it can only be used if the marginal models are all modeled parametrically.

To resolve this problem we will use the quantile regression approach, i.e., we will work with a different loss function instead of a likelihood. Quantile regression was introduced by Koenker & Bassett (1978); see also Koenker (2005). In contrast to classical regression, it provides a global picture of the predictors' effect on the distribution of the responses, while it is robust to heavy-tailed distributions. In high-dimensional settings Belloni & Chernozhukov (2011) studied linear quantile regression with the Lasso penalty, Wang et al. (2012) proved selection consistency of linear quantile regression with nonconvex penalty functions, and Sherwood & Wang (2016) derived asymptotic properties of partially linear additive quantile regression with a nonconvex penalty. In addition to these articles on single quantile regression, Zou & Yuan (2008a) introduced a composite quantile regression approach for linear models, which considers multiple quantiles simultaneously. They assumed that the slopes were the same across quantiles and used the adaptive Lasso penalty from Zou (2006). The method enjoys the oracle properties proposed in Fan & Li (2001). When heterogeneity exists, that is, when slopes vary across quantiles, the method of Zou & Yuan (2008b) is able to detect non-zero slopes simultaneously. Zou & Yuan (2008b) generalized the approach to the case with multiple responses. The two 2008 articles by Zou and Yuan consider only the scenario with a fixed number of parameters; Zou & Yuan (2008b) is a computational article without proofs of asymptotic properties.

The goal of Chapter 3 is simultaneous variable selection in multiple linear models with multiple quantiles. To take the unknown dependence structure between the responses in the different experiments into account, we integrate the data by summing up their quantile loss functions. Addi-

tionally we apply a nonconvex penalty on the L_1-norm of the coefficients related to each predictor, which represents overall strength of the predictor across multiple quantiles and experiments. The L_1-norm was also used in Sherwood & Wang (2016) for variable selection with multiple quantiles. In quantile regression settings, it enjoys computational convenience. Peng & Wang (2015) proposed a new "Quick Iterative Coordinate Descent" algorithm (QICD) for solving nonconvex penalized quantile regression in high-dimensions with no group structure. With slight modifications, the QICD algorithm can be easily adapted to our approach. Moreover, unlike the nonconvex penalty based on the L_2-norm used by Gao & Carroll (2017), which makes either all or none of parameters in a group zero, the L_1-norm allows a group to contain both zero and nonzero parameters (Jiang & Huang, 2015). It is reasonable because it is intuitively clear that a predictor, which significantly affects one of the responses, does not necessarily have effect on the response in a different experimental environment. Nevertheless, the natural grouping structure (the parameters of different quantiles and experiments that belong to one predictor) must be taken into account, i.e., the parameters of the same predictor should not be treated separately.

Multiple quantile regression with grouped nonconvex penalization for high-dimensional and dependent data from various experiments has, to the best of our knowledge, not been studied in the literature. We establish selection consistency and asymptotic normality of our estimator in this quite general setting under mild assumptions. Additionally we propose a multiple quantile Bayesian information criterion (MQBIC) based on pooled check functions, which is an extension of the Bayesian information criterion for linear quantile regression (Lee et al., 2014) to the multiple experiment scenario. Similar to the pseudolikelihood information criterion in Gao & Carroll (2017), MQBIC permits consistent model selection, see Section 3.2, and to choose the tuning parameter for the penalized estimator, see Section 3.3.

Chapter 3 is organized as follows. Section 3.1 introduces our objective function, which involves a nonconvex group penalization term, and present the oracle properties of the estimator. In Section 3.2, we propose the MQBIC and establishe its model selection consistency. Section 3.3 compares our method with other approaches using computer simulations. Our method is illustrated

in Section 3.4 by means of a real data example. Section 4.2 gives a brief conclusion of Chapter 3 and discussion for further questions. All the technical details are provided in Appendix B. For reasons of clarity we assume here that the sample sizes and the quantile levels are the same in every experiment. The conclusions obviously remain valid if we drop these assumptions.

2. EFFICIENT ESTIMATORS FOR EXPECTATIONS IN NONLINEAR PARAMETRIC REGRESSION MODELS WITH RESPONSES MISSING AT RANDOM

2.1 Expansion of the estimator

Our estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$ from (1.2) consists of the nonparametric estimator $\widehat{H}_{np}$ from equation (1.3) and a correction term $\widehat{\Gamma}$, which has the form

$$\widehat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} Z_i \widehat{g}(X_i) \widehat{\varepsilon}_i. \tag{2.1}$$

Here $\widehat{g}(x)$ is a consistent estimator of

$$g(x) = \frac{\rho_h(x)}{\pi(x)\sigma^2(x)}$$

uniformly in x over the support $\mathcal{I}$ of X, with $\rho_h(x) = E\{h(X,Y)\varepsilon | X = x\}$ and $\sigma^2(x) = E(\varepsilon^2 | X = x)$. The term $\widehat{\Gamma}$ incorporates the parametric regression structure and is suggested by the canonical gradient, which characterizes the influence function of an efficient estimator (see Section 2.2).

To estimate $g(x)$ we can, for example, use a combination of Nadaraya-Watson estimators introduced by Nadaraya (1964) and Watson (1964). The residuals $\widehat{\varepsilon}_i = Y_i - r_{\widehat{\vartheta}}(X_i)$ are based on an efficient estimator $\widehat{\vartheta}$ of ϑ; see Müller & Van Keilegom (2012) for an approach using estimating equations, and also for an overview of related efficient methods.

All estimators in $\widehat{\Gamma}$, including $\widehat{\vartheta}$, are complete case estimators since only observations with $Z = 1$ are used; see Müller & Schick (2017) which shows that in the model with MAR responses complete case analysis is efficient for estimating characteristics of the conditional distribution of Y given X. The estimator $\widehat{H}_{np}$ from (1.3), on the contrary, is an imputation estimator. Hence our estimator (1.2) is a combination of imputation and complete case analysis.

In the usual model with no missing data, the partially imputed estimator $\widehat{H}_{np}$ for $E\{h(X,Y)\}$

reduces to the empirical estimator. However, it is not efficient unless we enhance it by correcting for the unknown parametric regression function using $\widehat{\Gamma}$ with all $Z_i = 1$ ($i = 1, \ldots, n$) and $\pi \equiv 1$, i.e., the efficient estimator becomes

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} h(X_i, Y_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{\rho}_h(x)}{\widehat{\sigma}^2(x)}\widehat{\varepsilon}_i.$$

In the rest of this chapter and Appendix A, we will, for convenience of notations, always use the lower case letter c to represent a generic constant. The norm brackets $\|\cdot\|$ refer to the Euclidean norm of a vector.

We will assume throughout that $\pi(x) > 0$, for all x on the support $\mathcal{I}$ of X, to exclude the extreme case that no response is observed, that $h(X, Y)$ is square-integrable and that $E(\varepsilon^2)$ is positive and finite. The regression function needs to satisfy the following condition.

Assumption 2.1. *The regression function $\tau \mapsto r_\tau(x)$ is differentiable at $\tau = \vartheta$ with a p-dimensional square-integrable gradient $\dot{r}_\vartheta(x)$ that satisfies*

$$\sup_{x\in\mathcal{I}} \|\dot{r}_\tau(x) - \dot{r}_\vartheta(x)\| \leq L\,\|\tau - \vartheta\| \quad \textit{for some constant } L \in \mathbb{R}.$$

To construct an efficient estimator of $E\{h(X, Y)\}$, an efficient estimator of ϑ, say $\widehat{\vartheta}$, is needed. Efficient estimation of ϑ in models defined by conditional constraints is discussed in Müller & Van Keilegom (2012). They show that an efficient estimator $\widehat{\vartheta}$ is characterized by the following expansion.

Assumption 2.2. *The estimator $\widehat{\vartheta}$ of ϑ satisfies*

$$\widehat{\vartheta} - \vartheta = \frac{1}{n} I^{-1} \sum_{i=1}^{n} Z_i \dot{r}_\vartheta(X_i)\sigma^{-2}(X_i)\varepsilon_i + o_p(n^{-1/2}),$$

with $I = E\{Z\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\sigma^{-2}(X)\}$, which is assumed to be invertible.

An example of an efficient estimator is provided in Müller & Van Keilegom (2012), who pro-

pose using

$$\widehat{\vartheta} = \arg\min_{\theta} \Big\| \sum_{i=1}^{n} Z_i \dot{r}_\theta(X_i) \widehat{\sigma}^{-2}(X_i) \{Y_i - r_\theta(X_i)\} \Big\|,$$

where $\widehat{\sigma}^2(x)$ is a consistent estimator of $\sigma^2(x)$, uniformly in $x \in \mathcal{I}$, for example the Nadaraya-Watson estimator. Under the conditions of Theorem 2.1 in Müller & Van Keilegom (2012), the solution $\widehat{\vartheta}$ of the above estimating equation satisfies Assumption 2.2.

In the next two subsections we will expand the partially imputed estimator $\widehat{H}_{np}$ of equation (1.3) and derive the expansion of the correction term $\widehat{\Gamma}$ introduced in (2.1). Combining the two parts gives the expansion of $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$, which is stated in Corollary 2.1 at the end of this section.

2.1.1 Expansion of the nonparametric estimator

Consider the nonparametric partial imputation estimator introduced in (1.3), which imputes only the incomplete cases, as in Cheng (1994). We propose estimating the conditional expectation $\chi(x) = E\{h(X, Y)|X = x\}$ by the Nadaraya-Watson estimator

$$\widehat{\chi}(x) = \frac{\sum_{j=1}^{n} K_b(X_j, x) Z_j h(X_j, Y_j)}{\sum_{j=1}^{n} K_b(X_j, x) Z_j}$$

with $K_b(u, x) = b^{-d} K(b^{-1}(u - x), x)$, where $K(\cdot, x)$ is a kernel function with integrated boundary correction, i.e., the kernel's form is different for interior and boundary points x. The letter $b = b_n$ denotes a bandwidth sequence which tends to zero as n increases.

To derive the expansion of the partially imputed estimator (1.3) we stipulate the following assumptions on the covariate vector X and the kernel K.

Assumption 2.3. *The d-dimensional random vector X has a compact support $\mathcal{I}$ and a density f that is bounded and bounded away from zero on $\mathcal{I}$.*

Assumption 2.4. *The kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a function that satisfies the following properties.*

(i) The kernel K is bounded and

$$\int_{\mathbb{R}^d} \left| s_1^{q_1} ... s_d^{q_d} K(s,x) \right| ds < \infty$$

for $x \in \mathcal{I}$ and any non-negative integers $q_1, q_2, ..., q_d$ satisfying $q_1 + ... + q_d = d + 1$, where $s_1, ..., s_d$ are the components of s.

(ii) Denote the region $\mathcal{S}_b(x) = \{b^{-1}(y - x) : y \in \mathcal{I}\}$. Then

$$\int_{\mathcal{S}_b(x)} K(s,x) ds = 1 \text{ and } \int_{\mathcal{S}_b(x)} s_1^{l_1} ... s_d^{l_d} K(s,x) ds = 0$$

for $x \in \mathcal{I}$ and any non-negative integers $l_1, l_2, ..., l_d$ satisfying $0 < l_1 + ... + l_d < d + 1$.

(iii) $K(s,x)$ is differentiable with respect to s. For some constants η, ζ and $\nu > 1$, $\|\partial K(s,x)/\partial s\| \leq \eta$, and $\|\partial K(s,x)/\partial s\| \leq \eta \|s\|^{-\nu}$ for any s satisfying $\|s\| \geq \zeta$.

Assumption 2.4 is necessary because we consider a scenario with a covariate vector X. This is in contrast to Cheng (1994), who considers the nonparametric model with univariate covariates X. Cheng uses Theorem 1 of Devroye & Wagner (1980) to derive the expansion of his version of $\widehat{H}_{np}$. The theorem requires a non-negative kernel function, so it cannot be applied to our multivariate scenario, which requires using higher order kernels. For the construction of such kernels we can use results from Simonoff (2012); see Remark 2.1 at the end of this subsection for details.

For the ease of derivation we will further assume that $\pi(x)$ and $\sigma^2(x)$ are bounded away from zero on $\mathcal{I}$. In the second conclusion of Lemma A.1 below we will show that $\sum_{j=1}^n K_b(X_j, x) Z_j / n$ converges to $\pi(x) f(x)$ in probability, uniformly in x. It follows that

$$\left\{ \inf_{x \in \mathcal{I}} \left| \frac{1}{n} \sum_{j=1}^n Z_j K_b(X_j, x) \right| \right\}^{-1} < \infty \tag{2.2}$$

with probability tending to one. Hence we can suppose, without loss of generality, that the denominator in the Nadaraya-Watson estimator $\widehat{\chi}(\cdot)$ is bounded away from zero on $\mathcal{I}$. Finally we need

the following two conditions.

Assumption 2.5. *The bandwidth* $b = b_n$ *satisfies* $nb^{2d}(\log n)^{-2} \to \infty$ *and* $nb^{2(d+1)} \to 0$ *as* $n \to \infty$.

Assumption 2.6. *The functions* χ, π *and* f *are* $d+1$ *times continuously differentiable over* $\mathcal{I}$.

Based on the above four assumptions, Thereom 2.1 below gives the expansion of the nonparametric estimator $\widehat{H}_{np}$ in (1.3).

Theorem 2.1. *Suppose Assumptions 2.3, 2.4, 2.5 and 2.6 are satisfied. Then the nonparametric estimator* $\widehat{H}_{np}$ *given in (1.3) has the expansion*

$$\widehat{H}_{np} = \frac{1}{n}\sum_{i=1}^{n}\left[\chi(X_i) + \frac{Z_i}{\pi(X_i)}\{h(X_i, Y_i) - \chi(X_i)\}\right] + o_p(n^{-1/2}).$$

Remark 2.1. *To specify a kernel that satisfies Assumption 2.4, we can extend the construction of second order boundary kernels in Section 3.3.1 of Simonoff (2012) to higher order boundary kernels. Consider, for example, the case* $d = 2$ *and* $X = (X_1, X_2)^{\mathrm{T}}$, $s = (s_1, s_2)^{\mathrm{T}}$ *and* $x = (x_1, x_2)^{\mathrm{T}}$. *Based on four different univariate bounded functions* $L_i(\cdot),\ i = 1, \ldots, 4$, *which satisfy* $\int |s_1|^3 |L_i(s_1)| ds_1 < \infty$, $|\partial L_i(s_1)/\partial s_1| < \eta$, *and* $|\partial L_i(s_1)/\partial s_1| < \eta |s_1|^{\nu}$ *for any* s_1 *satisfying* $|s_1| > \zeta$, *with some constants* η, ζ *and* $\nu > 1$, *we first calculate second order kernels*

$$\begin{aligned} T_1(s_1, x_1) &= \frac{\ell_2^{(1)}(x_1) L_1(s_1) - \ell_1^{(1)}(x_1) L_2(s_1)}{\ell_2^{(1)}(x_1)\ell_1^{(0)}(x_1) - \ell_1^{(1)}(x_1)\ell_2^{(0)}(x_1)}, \\ T_2(s_1, x_1) &= \frac{\ell_4^{(1)}(x_1) L_3(s_1) - \ell_3^{(1)}(x_1) L_4(s_1)}{\ell_4^{(1)}(x_1)\ell_3^{(0)}(x_1) - \ell_3^{(1)}(x_1)\ell_4^{(0)}(x_1)}, \end{aligned}$$

with $\ell_i^{(j)}(x_1) = \int_{\mathcal{S}_{b,1}(x_1)} s_1^j L_i(s_1) ds_1$ *and* $S_{b,1}(x_1) = \{b^{-1}(y_1 - x_1) : y_1 \in \mathcal{I}_1\}$, *where* $\mathcal{I}_1$ *denotes the support of* X_1. *Then the linear combination of* $T_1(s_1, x_1)$ *and* $T_2(s_1, x_1)$,

$$K_1(s_1, x_1) = \frac{t_2^{(2)}(x_1) T_1(s_1, x_1) - t_1^{(2)}(x_1) T_2(s_1, x_1)}{t_2^{(2)}(x_1) - t_1^{(2)}(x_1)},$$

with $t_i^{(j)}(x_1) = \int_{\mathcal{S}_{b,1}(x_1)} s_1^j T_i(s_1, x_1) ds_1$*, is a univariate third order boundary kernel for* X_1*. A third order boundary kernel* $K_2(s_2, x_2)$ *for* X_2 *can be constructed analogously. By taking the product we obtain the desired bivariate third order boundary kernel* $K(s, x) = K_1(s_1, x_1)K_2(s_2, x_2)$ *for* X*.*

The construction of general multivariate higher order boundary kernels is done analogously. For $j = 2, 3 \ldots, d$*, we first calculate univariate* j^{th} *order boundary kernels* T_1 *and* T_2*, and then a univariate* $(j+1)$*-th order boundary kernel as the linear combination given above. The product of* j *such univariate* $(j+1)^{\text{th}}$ *boundary kernels yields a multivariate* $(j+1)^{\text{th}}$ *order boundary kernel* K*.*

A multivariate $(d+1)^{\text{th}}$ *order boundary kernel constructed in this way will satisfy Assumption 2.4. If the boundary of the support* $\mathcal{I}$ *is unknown, it can be estimated using extreme values, i.e.,* $(min_{1 \le i \le n}\{X_{i1}\}, \ldots, min_{1 \le i \le n}\{X_{id}\})^{\text{T}}$ *and* $(max_{1 \le i \le n}\{X_{i1}\}, \ldots, max_{1 \le i \le n}\{X_{id}\})^{\text{T}}$*.*

2.1.2 Expansion of the correction term

To expand the additive correction

$$\widehat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} Z_i \widehat{g}(X_i) \widehat{\varepsilon}_i,$$

the following assumption is required:

Assumption 2.7. *The function* $\rho_h(X) = E\{h(X, Y)\varepsilon | X\}$ *is square-integrable.*

Under Assumptions 2.1 on the regression function, as well as Assumption 2.2, 2.3 and 2.7, we expand the nonlinear correction $\widehat{\Gamma}$ in the next theorem. Remember that $g(x) = \rho_h(x)/\{\pi(x)\sigma^2(x)\}$.

Theorem 2.2. *Suppose that Assumptions 2.1, 2.2, 2.3 and 2.7 hold and that* $\widehat{g}(x)$ *is a consistent estimator of* $g(x)$*, uniformly in* $x \in \mathcal{I}$*. Then the nonlinear correction* $\widehat{\Gamma} = \sum_{i=1}^{n} Z_i \widehat{g}(X_i)\widehat{\varepsilon}_i/n$ *has the expansion*

$$\widehat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \frac{Z_i \varepsilon_i}{\sigma^2(X_i)} \Big\{ \frac{\rho_h(X_i)}{\pi(X_i)} - \dot{r}_\vartheta(X_i)^{\text{T}} I^{-1} \Delta \Big\} + o_p(n^{-1/2})$$

with $\Delta = E\{Z\dot{r}_\vartheta(X)g(X)\} = E\{\dot{r}_\vartheta(X)h(X,Y)\sigma^{-2}(X)\varepsilon\}$.

A common estimator of $g(x)$ is $\widehat{g}(x) = \widehat{\rho}_h(x)/\{\widehat{\sigma}^2(x)\widehat{\pi}(x)\}$ with $\widehat{\rho}_h(x)$, $\widehat{\sigma}^2(x)$ and $\widehat{\pi}(x)$ being Nadaraya-Watson estimators of $\rho_h(x)$, $\sigma^2(x)$ and $\pi(x)$, respectively. The Nadaraya-Watson estimator is uniformly consistent when X has a compact support. In Section 3.3 we will use this estimator for our simulation study, and also show more details.

We conclude the section with the final expansion of our estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$. The result follows directly from the statements in Theorems 2.1 and 2.2 on $\widehat{H}_{np}$ and $\widehat{\Gamma}$. We therefore formulate the result as a corollary.

Corollary 2.1. *Write* $\Delta = E\{\dot{r}_\vartheta(X)h(X,Y)\sigma^{-2}(X)\varepsilon\}$ *as in Theorem 2.2 and let the assumptions of that theorem be satisfied. Suppose that Assumptions 2.3, 2.4, 2.5 and 2.6 from Section 2.1.1 hold true. Then the estimator* $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$ *from equation (1.2) has the expansion*

$$
\begin{aligned}
&n^{1/2}[\widehat{H} - E\{h(X,Y)\}] \\
&= n^{-1/2}\sum_{i=1}^{n}\Big[\chi(X_i) - E\{h(X,Y)\} + \frac{Z_i}{\pi(X_i)}\{h(X_i,Y_i) - \chi(X_i)\} \\
&\qquad\qquad - \frac{Z_i\varepsilon_i}{\sigma^2(X_i)}\Big\{\frac{\rho_h(X_i)}{\pi(X_i)} - \dot{r}_\vartheta(X_i)^{\mathrm{T}} I^{-1}\Delta\Big\}\Big] + o_p(n^{-1/2}).
\end{aligned}
$$

2.2 Efficiency

In this section we calculate the canonical gradient of $E\{h(X,Y)\}$, which characterizes the influence function of an efficient estimator of that expectation. The efficiency of our estimator will be established by showing that the canonical gradient equals the influence function obtained in Section 2.1. We will use results from Müller et al. (2006) and Müller (2009) about the canonical gradient, and also from Schick (1993) about the tangent space in nonlinear regression.

Essential for the derivation of canonical gradients is the notion of tangent space: a canonical gradient is characterized as an orthogonal projection of a gradient onto the tangent space, which is the closed linear span of the set of all perturbations of the joint distribution $P(dx, dy, dz)$ within the model. The distribution P depends on the marginal distribution $G(dx)$ of X, the conditional

probability $\pi(x)$ of $Z = 1$ given $X = x$ and the conditional distribution $Q(x, dy)$ of Y given $X = x$. Müller et al. (2006), who also consider regression models with MAR responses, were the first to describe the tangent space for general differentiable functions $\kappa(G, Q, \pi)$ in this model. They write the joint distribution in the form

$$P(dx, dy, dz) = G(dx)B_{\pi(x)}\{zQ(x, dy) + (1 - z)\delta_0(dy)\},$$

with $B_p = p\delta_1 + (1 - p)\delta_0$ denoting the Bernoulli distribution with parameter p and δ_t the Dirac measure at t. To specify the tangent space we assume that G, Q and π have Hellinger differentiable perturbations:

$$\begin{aligned} G_{nu}(dx) &\doteq G(dx)\{1 + n^{-1/2}u(x)\}, \\ Q_{nv}(x, dy) &\doteq Q(x, dy)\{1 + n^{-1/2}v(x, y)\}, \\ B_{\pi_{nw}(x)}(dz) &\doteq B_{\pi(x)}(dz)[1 + n^{-1/2}\{z - \pi(x)w(x)\}], \end{aligned} \tag{2.3}$$

where $\doteq$ means ignoring $o_p(n^{-1/2})$ items. Since the perturbed distributions are probability distributions, the Hellinger derivative u belongs to

$$L_{2,0}(G) = \Big\{u \in L_2(G) : \int u dG = 0\Big\},$$

v belongs to

$$V_0 = \Big\{v \in L_2(M) : \int v(x, y)Q(x, dy) = 0\Big\},$$

with $M(dx, dy) = Q(x, dy)G(dx)$, and w belongs to

$$W = \big\{w \in L_2(G_\pi) : G_\pi(dx) = \pi(x)\{1 - \pi(x)\}G(dx)\big\}.$$

The tangent space is the orthogonal sum

$$\{u(X); u \in U\} \oplus \{Zv(X,Y) : v \in V\} \oplus \{\{Z - \pi(X)\}w(X) : w \in W\}.$$

As in Müller et al. (2006), we have no structural assumptions on G and π. This means that we have no further restrictions on the perturbations u and w and can therefore take $u \in U = L_{2,0}(G)$ and $w \in W = L_2(G_\pi)$. We must, however, take the regression structure into account, i.e. the space V is the subset of V_0 to which v is now restricted. In the following we assume that the subspaces U, V and W are closed and linear.

The canonical gradient g_* is an element of the tangent space and has the form

$$g_*(X, ZY, Z) = u_*(X) + Zv_*(X,Y) + \{Z - \pi(X)\}w_*(X), \tag{2.4}$$

where $u_*(X)$, $Zv_*(X)$ and $\{Z - \pi(X)\}w_*(X)$ are projections of the gradient (that characterizes the differentiable functional) onto the three orthogonal subspaces of the tangent space.

Until here we have only summarized results from Müller et al. (2006), who provide a detailed characterization of efficient estimators in the model with MAR responses on pages 352-355 and then specialize them to four specific models for the conditional distribution Q. In this paper we have $Q(x, dy) = f\{y - r_\vartheta(x)|x\}dy$, where $f(\cdot|x)$ denotes the conditional density of the (conditional mean zero) error distribution given $X = x$. In order to find V we introduce perturbations of the parameter ϑ and the conditional error density. The exact form of V and the derivation of u_*, v_* and w_* are located in the proof of Theorem 2.3; see Section A.4 for details. In the next theorem we provide the explicit representation of the canonical gradient of $E\{h(X,Y)\}$. The efficiency of our estimator is formulated subsequently in Corollary 2.2

Theorem 2.3. *Let the vector Δ and the matrix I be defined as in the previous section, i.e. $\Delta = E\{\dot{r}_\vartheta(X)h(X,Y)\sigma^{-2}(X)\varepsilon\}$ and $I = E\{Z\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\sigma^{-2}(X)\}$. Suppose Assumptions 2.1, 2.7 and the Hellinger differentiability assumption (2.3) are satisfied. Futher assume that the conditional density of ε given x, $f(\cdot|x)$, has a finite Fisher information I and that I is invertible. Then*

the canonical gradient of the functional $E\{h(X,Y)\}$ is

$$g_*(X, ZY, Z) = \chi(X) - E\{h(X,Y)\} + \frac{Z}{\pi(X)}\{h(X,Y) - \chi(X)\} - \frac{Z\varepsilon}{\sigma^2(X)}\Big\{\frac{\rho_h(X)}{\pi(X)} - \dot{r}_\vartheta(X)^{\mathrm{T}} I^{-1}\Delta\Big\}.$$

It follows from Corollary 2.1 that our estimator is asymptotically linear, and from Theorem 2.3 that the influence function given in Corollary 2.1 equals the canonical gradient g_* from Theorem 2.3. Hence our estimator is efficient in the sense of the Hájek and Le Cam theory, which implies asymptotic normality. We formulate this as a corollary.

Corollary 2.2. *Let the assumptions of Corollary 2.1 and Theorem 2.3 be satisfied. Then the estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$ introduced in equation (1.2) is asymptotically efficient. In particular it is asymptotically normally distributed with variance $E\{g_*(X, ZY, Z)^2\}$, with $g_*(X, ZY, Z)$ specified in Theorem 2.3 above.*

2.3 Simulations

2.3.1 Linear and nonlinear regression with one covariate

To illustrate the results of the previous sections, we conduct a simulation study comparing various estimators for $E(Y)$, $E(Y^2)$, $E(XY)$ and $E\{X\exp(XY)\}$; see Tables 2.1-2.4. In each case we consider two regression functions, $r_\vartheta(x) = \vartheta x$ and $r_\vartheta(x) = \cos(\vartheta x)$ with $\vartheta = 2$, and two variance functions, namely a linear variance function $\sigma^2(x) = 0.6 - 0.5x$ and a parabolic variance function $\sigma^2(x) = (x - 0.4)^2 + 0.1$. The covariate X is generated from a uniform distribution on $[-1, 1]$ and the error variable η in $\varepsilon = \sigma(X)\eta$ from a standard normal distribution. In all scenarios we use the logistic distribution function $\pi(x) = 1/\{1 + \exp(-x)\}$ for the conditional probability, so that about half of the simulated responses are missing. In this section we use, for simplicity, only ordinary kernels instead of the boundary kernels discussed in Remark 2.1.

To evaluate the performance of our asymptotically optimal estimator when sample sizes are small we simulate the mean squared errors (MSE) of $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$. Here $\widehat{H}_\vartheta$ denotes the version of

our estimator $\widehat{H}$ from (1.2) that uses the true values of $\sigma^2(x)$, $\pi(x)$ and ϑ in the correction term, whereas $\widehat{H}_{\widehat{\vartheta}}$ uses estimators for those quantities. For the calculation of

$$\widehat{\vartheta} = \arg\min_{\theta} \Big| \sum_{i=1}^{n} Z_i \widehat{\sigma}^{-2}(X_i) \dot{r}_\theta(X_i) \{Y_i - r_\theta(X_i)\} \Big|$$

we use a consistent nonparametric estimator for $\sigma^2(x)$, namely

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} Z_i K_{b_1}(x - X_i) \{Y_i - r_{\widehat{\vartheta}_0}(X_i)\}^2}{\sum_{i=1}^{n} Z_i K_{b_1}(x - X_i)},$$

where $K_{b_1}(\cdot)$ is a Gaussian kernel with bandwidth b_1 and $\widehat{\vartheta}_0$ is the ordinary least squares estimator (or some other consistent estimator of ϑ). In the model with a linear regression function $\widehat{\vartheta}$ and $\widehat{\vartheta}_0$ have a closed form, while for the cosine function we use the "`nls`" function in `R` to obtain them. Our nonparametric estimator for $\pi(x)$ is

$$\widehat{\pi}(x) = \frac{\sum_{i=1}^{n} Z_i K_{b_2}(x - X_i)}{\sum_{i=1}^{n} K_{b_2}(x - X_i)}, \tag{2.5}$$

where $K_{b_2}(\cdot)$ is a Gaussian kernel with bandwidth b_2; $\widehat{\rho}_h(x)$ is a plug-in estimator for $\rho_h(x) = E\{h(X,Y)\varepsilon | X = x\}$. For our choices of h it will involve the estimators $\widehat{\sigma}^2(x)$ and $\widehat{\vartheta}$ just described; see below for more details. The nonparametric part $\widehat{H}_{np}$ of our estimator $\widehat{H}$ is the partially imputed estimator (1.3). It is based on a Nadaraya-Watson estimator for the conditional expectation $\chi(x) = E\{h(X,Y)|X = x\}$, with a Gaussian kernel $K_{b_3}(\cdot)$ with bandwidth b_3.

We also compare $\widehat{H}_{\vartheta}$ and $\widehat{H}_{\widehat{\vartheta}}$ with $S = n^{-1} \sum_{i=1}^{n} \{Z_i h(X_i, Y_i)/\pi(X_i)\}$, the simple Horvitz-Thompson type estimator based on the true $\pi(x)$, and the nonparametric estimator $\widehat{H}_{np}$ without the nonlinear correction. For each setting simulations with sample sizes $n = 50$, 100 and 200 are conducted based on $5,000$ repetitions. The "nls" routine does not always converge for the cosine regression function. We therefore list the MSEs of $\widehat{H}_{\widehat{\vartheta}}$ only for sample sizes $n = 100$ and $n = 200$ for that scenario.

For estimators involving kernel estimation we use leave-one-out cross validation to select the

bandwidth. For example, to obtain the bandwidth b_1 of $\widehat{\sigma}^2(x)$, we first calculate, for each complete observation (X_j, Y_j),

$$\widetilde{\sigma}^2_{jb} = \sum_{\substack{i=1 \\ i\neq j}}^{n} Z_i K_b(X_j - X_i)\{Y_i - r_{\widehat{\vartheta}_0}(X_i)\}^2 \Big/ \sum_{\substack{i=1 \\ i\neq j}}^{n} Z_i K_b(X_j - X_i),$$

for bandwidths b from a candidate set $\mathcal{G}$. Then b_1 is obtained as

$$b_1 = \arg\min_{b\in\mathcal{G}} \sum_{i=1}^{n} Z_i[\widetilde{\sigma}^2_{ib} - \{Y_i - r_{\widehat{\vartheta}_0}(X_i)\}^2]^2.$$

For the case $h(x, y) = y$, for example, we used the set $\mathcal{G} = \{0.1, 0.2, ...0.5\}$ for b_1 and also for b_2. The bandwidth b_3 for the nonparametric part $\widehat{H}_{np}$ has the form $b_3 = an^{-2/5}$, which is indicated to have optimal convergence rate by Cheng (1994). We chose $a = 0.5, 0.6, \ldots, 0.9$ to determine b_3.

Table 2.1: Simulated mean square error of estimators of $E(Y)$

		$r_\vartheta(X) = \vartheta X \ \ (\vartheta = 2)$				$r_\vartheta(X) = \cos(\vartheta X) \ \ (\vartheta = 2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
	50	0.0341	0.0319	0.0634	0.0941	0.0102	-	0.0390	0.0426
(a)	100	0.0152	0.0144	0.0291	0.0443	0.0037	0.0071	0.0179	0.0215
	200	0.0070	0.0067	0.0144	0.0229	0.0015	0.0031	0.0085	0.0104
	50	0.0353	0.0323	0.0655	0.0968	0.0112	-	0.0411	0.0451
(b)	100	0.0157	0.0146	0.0308	0.0462	0.0041	0.0082	0.0195	0.0233
	200	0.0072	0.0068	0.0151	0.0236	0.0016	0.0035	0.0092	0.0112

The entries in both the left and the right panel are the simulated mean squared errors of estimators of the mean response. The first two columns of each panel show the MSEs of the two versions $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ of the efficient estimator. The third and fourth column list the results for the nonparametric estimator $\widehat{H}_{np}$ (no correction) and the simple estimator S. The variance functions are (a) $\sigma^2(X) = 0.6 - 0.5X$ and (b) $\sigma^2(X) = (X - 0.4)^2 + 0.1$.

The simulated mean squared errors for estimating the mean response are given in Table 2.1. In

this case $\rho_h(x) = E\{h(X,Y)\varepsilon|X = x\} = E\{Y\varepsilon|X = x\} = \sigma^2(x)$. In each row of Table 2.1 the efficient estimator outperforms the nonparametric estimator without the nonlinear correction, while the simple estimator is inferior to any of its competitors. In the linear regression model the two versions $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ of the efficient estimator differ slightly, in contrast to the cosine regression model, where the difference is quite large. This is because the estimator of the slope in the linear regression model is better than that of the frequency parameter in the model with the cosine regression function. The MSEs for different sample sizes confirm the root-n convergence rate of the efficient estimator, as stated in Corollary 2.1.

Table 2.2: Simulated mean square error of estimators of $E(Y^2)$

		$r_\vartheta(X) = \vartheta X \ \ (\vartheta = 2)$				$r_\vartheta(X) = \cos(\vartheta X) \ \ (\vartheta = 2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
	50	0.1235	0.1725	0.2755	0.4206	0.0626	-	0.1065	0.1267
(a)	100	0.0545	0.0818	0.1394	0.2099	0.0285	0.0275	0.0493	0.0630
	200	0.0247	0.0381	0.0660	0.1029	0.0134	0.0133	0.0234	0.0307
	50	0.1973	0.2484	0.4135	0.6012	0.0924	-	0.1207	0.1520
(b)	100	0.0891	0.1180	0.2130	0.3060	0.0456	0.0448	0.0592	0.0786
	200	0.0402	0.0544	0.1010	0.1475	0.0215	0.0214	0.0281	0.0375

The entries are mean squared errors as in Table 2.1, now with $h(x, y) = y^2$. The variance functions are again (a) $\sigma^2(X) = 0.6 - 0.5X$ and (b) $\sigma^2(X) = (X - 0.4)^2 + 0.1$.

Table 2.2 displays the simulation results for the same scenario as in Table 2.1, but now the second moment of the response is estimated. The efficient estimator again outperforms both the nonparametric estimator and the simple estimator. In our scenario with normal errors we have $\rho_h(x) = 2r_\vartheta(x)\sigma^2(x)$ with $r_\vartheta(x) = \vartheta x$ and $r_\vartheta(x) = \cos(\vartheta x)$. In both regression models it does not make a big difference whether true values or estimators are used.

The MSEs for estimating $E(XY)$ are given in Table 2.3. In both regression models $\rho_h(x) =$

Table 2.3: Simulated mean square error of estimators of $E(XY)$

		$r_\vartheta(X) = \vartheta X \ \ (\vartheta = 2)$				$r_\vartheta(X) = \cos(\vartheta X) \ \ (\vartheta = 2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
	50	0.0120	0.0147	0.0215	0.0402	0.0034	-	0.0136	0.0131
(a)	100	0.0050	0.0068	0.0105	0.0203	0.0013	0.0009	0.0065	0.0068
	200	0.0022	0.0030	0.0048	0.0099	0.0005	0.0004	0.0031	0.0032
	50	0.0131	0.0159	0.0268	0.0456	0.0044	-	0.0186	0.0187
(b)	100	0.0055	0.0073	0.0134	0.0231	0.0017	0.0011	0.0093	0.0096
	200	0.0023	0.0032	0.0062	0.0113	0.0007	0.0004	0.0044	0.0046

We consider the same scenario as in Tables 2.1 and 2.2, now with $h(x, y) = xy$.

$x\sigma^2(x)$. The first two columns of the left panel (linear regression) indicate that estimating $\sigma^2(x)$, $\pi(x)$ and ϑ increases the MSE slightly. The MSEs in the corresponding columns in the right panel (cosine regression) appear to be similar. The results in Table 2.3 again confirm the superiority of the efficient estimator as well as the convergence rate.

The results for $E\{X \exp(XY)\}$ are listed in Table 2.4. Straightforward calculations yield $\rho_h(x) = \sigma^2(x)x^2 \exp[\{\vartheta + \sigma^2(x)/2\}x^2]$. The efficient estimator clearly outperforms the competing estimators. As in the previous tables we see that the two estimators $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ based on true values and on estimates perform similarly.

The influence function of the efficient estimator in Corollary 2.1 contains a non-negligible part that comes from the difference $n^{1/2}(\widehat{\vartheta} - \vartheta)$. This part is missing if we replace $\widehat{\vartheta}$ by ϑ, which explains why in some cases, e.g., the upper left panel in Table 2.4, $\widehat{H}_{\widehat{\vartheta}}$ outperforms $\widehat{H}_\vartheta$. However, estimating $\sigma^2(x)$ and $\pi(x)$ adds uncertainty, especially if n is not very large, so that in other cases, for example in the right panel in Table 2.4, the MSE of $\widehat{H}_{\widehat{\vartheta}}$ can be larger than that of $\widehat{H}_\vartheta$.

Table 2.4: Simulated mean square errors of estimators of $E\{X\exp(XY)\}$

		$r_\vartheta(X)=\vartheta X\ \ (\vartheta=2)$				$r_\vartheta(X)=\cos(\vartheta X)\ \ (\vartheta=2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
	50	0.5273	0.4491	0.7517	1.2958	0.0286	-	0.0415	0.0727
(a)	100	0.2442	0.2164	0.3693	0.6207	0.0136	0.0144	0.0210	0.0350
	200	0.1289	0.1197	0.2161	0.3420	0.0072	0.0075	0.0122	0.0191
	50	2.2743	1.9389	3.0689	5.0159	0.1148	-	0.1566	0.2547
(b)	100	0.9657	0.8976	1.3701	2.2641	0.0491	0.0504	0.0706	0.1166
	200	0.4589	0.4624	0.7310	1.4083	0.0238	0.0264	0.0383	0.0710

In this table $h(x,y)=x\exp(xy)$; the scenario is the same as in Tables 2.1-2.3.

2.3.2 Linear regression with two covariates

Finally we consider a bivariate covariate vector $X=(X_1,X_2)^{\mathrm{T}}$ and a linear regression function $r_\vartheta(x)=\vartheta_1x_1+\vartheta_2x_2$ with $\vartheta_1=1$ and $\vartheta_2=2$. We modify the scenario of the previous section as follows: the variance function $\sigma^2(x)=\sigma^2(x_1,x_2)$ is set to be $2.1-0.5(x_1+x_2)$ or $(x_1+x_2-0.8)^2+0.1$, and $\pi(x)=1/[1+\exp\{-(x_1+x_2)\}]$. In order to generate correlated covariates X_1,X_2 we first sample auxiliary random variables W,X_1' and X_2' independently: W is generated from a uniform distribution on $[-0.5,0.5]$ and X_1' and X_2' from a uniform distribution on $[-1,1]$. Then we take $X_1=X_1'+W$ and $X_2=X_2'+W$. Our final estimator is based on kernel estimators. For example, $\widehat{\sigma}^2(x)$ now involves a product of two Gaussian-based kernels of order 4 (Wand & Schucany, 1990), i.e., $K(x)=(3-x^2)\Phi(x)/2$ where $\Phi(\cdot)$ is the standard Gaussian density function, both using the same bandwidth, to estimate the unknown conditional expectations. Table 2.5 shows the simulated mean squared errors of estimators of the mean response in the bivariate regression model. In this case $\rho_h(x)=E\{h(X,Y)\varepsilon|X=x\}=\sigma^2(x)$. Again our efficient estimator outperforms the competing estimators and confirms our theoretical results. The efficient estimator that uses estimates $\widehat{\sigma}^2(x)$, $\widehat{\pi}(x)$ and $\widehat{\vartheta}$ is better than the estimator $\widehat{H}_\vartheta$, which uses the true values.

Table 2.5: Simulated mean square error of estimators of $E(Y)$

$\sigma^2(X)$	n	$\widehat{H}_{\vartheta}$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
$2.1 - 0.5(X_1 + X_2)$	50	0.1081	0.0963	0.2022	0.2991
	100	0.0514	0.0444	0.1013	0.1529
	200	0.0247	0.0214	0.0507	0.0759
$(X_1 + X_2 - 0.8)^2 + 0.1$	50	0.1263	0.1054	0.1876	0.3338
	100	0.0617	0.0514	0.1020	0.1738
	200	0.0297	0.0248	0.0550	0.0861

The entries are simulated mean squared errors of estimators of the mean response, now for the scenario with the bivariate linear regression function $r_{\vartheta}(X) = \vartheta_1 X_1 + \vartheta_2 X_2$ ($\vartheta_1 = 1$, $\vartheta_2 = 2$) described in Section 2.3.2.

2.4 Real data analysis

In this section we apply our method to a data set of 2139 HIV positive patients from a clinical trial (Hammer et al., 1996). The data are freely accessible in the `R` package `speff2trial`.

In the trial the patients were randomly assigned to four antiretroviral therapies: (i) zidovudine (ZDV) monotherapy, (ii) ZDV + didanosine (DDI), (iii) ZDV + zalcitabine, and (iv) DDI monotherapy. We want to compare the ZDV monotherapy (i) with the alternative group of therapies (ii)-(iv), and estimate the mean number of CD4 cells in both groups, i.e. the number of white blood cells that fight the infection. An increasing CD4 count indicates that the HIV treatment is more effective.

We are interested in the difference between the mean CD4 counts (Y) in the monotherapy group and the mean CD4 counts in the alternative therapy group at 96$\pm$5 weeks post therapy. There are six covariates: $X^{(1)}$, age; $X^{(2)}$, weight; $X^{(3)}$, CD4 counts at baseline; $X^{(4)}$, CD4 counts at 20$\pm$5 weeks; $X^{(5)}$, CD8 (immune cells) counts at baseline; $X^{(6)}$, CD8 counts at 20$\pm$5 weeks. Because of death and dropout, 39% of the responses in the monotherapy group and 37% of the responses of the combined therapy group are missing, while all covariates are observed for all patients. Let Z again

denote the missingness indicator (which is 1 if Y is observed and 0 if it is missing). As indicated by Hu et al. (2010) and Tang et al. (2018), who consider the same data set, it is reasonable to asume that the conditional expectation of the response given the covariates can be modelled using linear regression, and that the response is missing at random. The variable selection results in Tang et al. (2018) suggest that only $X^{(3)}$, $X^{(4)}$ and $X^{(6)}$ actually affect Y. We therefore assume $E(Y|X) = \vartheta^{\mathrm{T}} X$, with a covariate vector $X = (1, X^{(3)}, X^{(4)}, X^{(6)})^{\mathrm{T}}$ and a regression parameter $\vartheta \in \mathbb{R}^4$.

We apply our method to the two groups of data separately and construct the efficient estimator for the mean response $\mu^{(0)}$ in the monotherapy group and the mean response $\mu^{(1)}$ in the combined therapy group. Then we calculate the difference between the means, $\mu^{(1)} - \mu^{(0)}$. For the construction of the efficient estimator see Section 2.3.1.

For comparison we also consider the three estimators for the mean difference $\mu^{(1)} - \mu^{(0)}$ in Section 7 of Hu et al. (2010): inverse probability weighting estimation, augmented inverse probability weighting estimation, and semiparametric dimension reduction estimation. In addition to the linear regression model between Y and X, Hu et al. (2010) additionally assume a parametric logistic model for the probablity of missingness, i.e. $\mathrm{logit}\{\pi(X)\} = \gamma^{\mathrm{T}} X$ for some parameter γ (which is technically a different model). For the term $\pi(X)$ in the nonlinear correction term of our efficient estimator, we therefore use the nonparametric estimator (2.5) and a parametric estimator for the logistic model, both based on $(X^{(3)}, X^{(4)}, X^{(6)})^{\mathrm{T}}$.

The point estimators, standard errors and 95% confidence intervals of various methods are given in Table 2.6. The results of the IPW, AIPW and SDR are taken from Hu et al. (2010) for comparison. The standard errors of the EEP and EENP are obtained using the bootstrap based on 500 repetitions. The point estimators, standard errors and confidence intervals of our method are close to those of the AIPW and SDR, which both attain an efficiency bound if $E(Y|X)$ and $\pi(X)$ are correctly specified, as discussed in Section 3 of Hu et al. (2010). However, our method is efficient without specifying an auxiliary parametric model of $\pi(X)$. From Table 2.6 we can see that the results of the two approaches (with and without a parametric) are very close.

Table 2.6: Estimators of the mean difference $\mu^{(1)} - \mu^{(0)}$

	Point estimator	Standard error	95% confidence interval
IPW	58.19	10.33	[37.94, 78.44]
AIPW	61.91	8.83	[44.60, 79.22]
SDR	62.42	9.02	[44.74, 80.10]
EENP	63.75	9.07	[45.98, 81.52]
EEP	63.40	9.08	[45.60, 81.20]

Here IPW, inverse probability weighting estimation; AIPW, augmented inverse probability weighting estimation; SDR, semiparametric dimension reduction estimation; EENP, efficient estimator with the nonparametric estimator for the probability of missingness; EEP, a version of the efficient estimator with the logistic model for the probability of missingness.

3.1 Penalized estimator

Throughout this chapter and Appendix B we will use the lower case letter c to represent a generic constant and I_m to mean a $m \times m$ identity matrix. The notation $\|\cdot\|_1$, $\|\cdot\|$ and $\otimes$ refer to the L_1- and L_2- norm of a vector and the Kronecker product, respectively.

Our quantile regression model is given by the constraint

$$Q_{\tau_m}(Y_k|X_k) = X_k^{\mathrm{T}}\theta_{km}^*,$$

where $Q_{\tau_m}(Y_k|X_k)$ is the $\tau_m \times 100\%$ conditional quantile of Y_k given X_k for $m = 1, \ldots, M$, and $0 < \tau_1 < \tau_2 < \cdots < \tau_M < 1$. This holds, for example, if the random error and the predictors in (1.4) are dependent, that is, $U_k = (X_k^{\mathrm{T}}\gamma_k^*)\xi_k$ for some $p_n \times 1$ vector γ_k^* and random variable ξ_k independent of X_k. We can set the first column of X_k to be $(1, \ldots, 1)^{\mathrm{T}}$ so that the model contains intercept terms. The number of predictors p_n tends to infinity as the sample size n increases.

For each $k = 1, \ldots, K$, we consider n independent copies $\{Y_{ki}, X_{ki} = (X_{ki1}, \ldots, X_{kip_n})^{\mathrm{T}}\}$, $i = 1, \ldots, n$, of the base observation $\{Y_k, X_k\}$ from model (1.4). Here we use three subscripts to locate the predictors, i.e., X_{kij} represents the j^{th} component of the i^{th} observation in the k^{th} experiment. We write $X_{k\cdot j} = (X_{k1j}, \ldots, X_{knj})^{\mathrm{T}}$ for the vector. The data are summarized in Table 3.1.

The regression parameters θ_{km}^* is assumed to be sparse for $k = 1, \ldots, K$ and $m = 1, \ldots, M$, that is, most of their components are zero. For $j = 1, \ldots, p_n$, define the parameters related to the j^{th} predictors across the experiments and quantile levels as

$$\theta^{*(j)} = (\theta_{11j}^*, \ldots, \theta_{1Mj}^*, \ldots, \theta_{K1j}^*, \ldots, \theta_{KMj}^*)^{\mathrm{T}}.$$

We want to select the predictors that have influence on any of the responses, i.e., we want to specify

Table 3.1: Data structure of multiple experiments

	Experiment 1	...	Experiment K
Parameters of τ_1	$\theta^*_{11} = (\theta^*_{111}, \ldots, \theta^*_{11p_n})^{\mathrm{T}}$	...	$\theta^*_{K1} = (\theta^*_{K11}, \ldots, \theta^*_{K1p_n})^{\mathrm{T}}$
$\vdots$	$\vdots$		$\vdots$
Parameters of τ_M	$\theta^*_{1M} = (\theta^*_{1M1}, \ldots, \theta^*_{1Mp_n})^{\mathrm{T}}$	...	$\theta^*_{KM} = (\theta^*_{KM1}, \ldots, \theta^*_{KMp_n})^{\mathrm{T}}$
Observation 1	Y_{11}, $X_{11} = (X_{111}, \ldots, X_{11p_n})^{\mathrm{T}}$	...	Y_{K1}, $X_{K1} = (X_{K11}, \ldots, X_{K1p_n})^{\mathrm{T}}$
$\vdots$	$\vdots$		$\vdots$
Observation n	Y_{1n}, $X_{1n} = (X_{1n1}, \ldots, X_{1np_n})^{\mathrm{T}}$	...	Y_{Kn}, $X_{Kn} = (X_{Kn1}, \ldots, X_{Knp_n})^{\mathrm{T}}$

the set $\mathcal{A} = \{j : 1 \leq j \leq p_n, \|\theta^{*(j)}\| > 0\}$. Without loss of generality let $\mathcal{A} = \{1, 2, \ldots, q_n\}$, i.e., only the first q_n predictors have nonzero parameters. We assume that q_n tends to infinity as n increases. For convenience, when the last subscript of the parameter or predictor is a, this refers to subvectors or submatrices consisting of only components with subscripts in $\mathcal{A}$. For example, $X_{kia} = (X_{ki1}, \ldots, X_{kiq_n})^{\mathrm{T}}$, $X_{k\cdot a} = (X_{k1a}, \ldots, X_{kna})^{\mathrm{T}}$ and $\theta^*_{kma} = (\theta^*_{km1}, \ldots, \theta^*_{kmq_n})^{\mathrm{T}}$.

The dependence between the experiments is unspecified. To integrate their data we therefore sum up their quantile loss functions across M different quantiles, i.e.,

$$\ell_n(\theta) = n^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki}^{\mathrm{T}} \theta_{km}) \tag{3.1}$$

where $\rho_m(x) = x\{\tau_m - I(x < 0)\}$ is the check function, $I(\cdot)$ is a 0-1 indicator function, and $\theta = (\theta_{11}^{\mathrm{T}}, \ldots, \theta_{1M}^{\mathrm{T}}, \ldots, \theta_{K1}^{\mathrm{T}}, \ldots, \theta_{KM}^{\mathrm{T}})^{\mathrm{T}}$ for $k = 1, \ldots, K$. To select the predictors that affect any of the responses, a penalty function $\Omega_{\lambda_n}(\cdot)$, with a tuning parameter λ_n, is imposed on the overall impact of each predictor. That impact is represented by the L_1 norm of the vector $\theta^{(j)}$, i.e. of its parameters across the K experiments. This gives the overall objective function

$$\Gamma_{\lambda_n}(\theta) = \ell_n(\theta) + \sum_{j=1}^{p_n} \Omega_{\lambda_n}(\|\theta^{(j)}\|_1). \tag{3.2}$$

Our estimator is obtained by minimizing $\Gamma_{\lambda_n}(\theta)$. We use the smoothly clipped absolute deviation (SCAD) penalty function (Fan & Li, 2001)

$$\Omega_{\lambda_n}(x) \quad = \quad \lambda_n x I(0 \leq x \leq \lambda_n) + \frac{a\lambda_n x - (x^2 + \lambda_n^2)/2}{a-1} I(\lambda_n < x < a\lambda_n) + \frac{(a+1)\lambda_n^2}{2} I(x \geq a\lambda_n)$$

where a is a constant that is usually set to 3.7 (Fan & Li, 2001). Before stating the asymptotic properties of our estimator, we make the following assumptions:

Assumption 3.1. *There is a constant $M_1 > 0$ such that $|X_{kij}| \leq M_1$ for every k, i and j.*

Assumption 3.2. *For every k there are constants $0 < M_2 \leq M_3$ such that*

$$M_2 \leq \lambda_{\min}(n^{-1} X_{k\cdot a}^{\mathrm{T}} X_{k\cdot a}) \leq \lambda_{\max}(n^{-1} X_{k\cdot a}^{\mathrm{T}} X_{k\cdot a}) \leq M_3$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ stand for the smallest and the largest eigenvalue, respectively. In addition, $X_{k\cdot a}$ and $(Y_{k1}, \ldots, Y_{kn})^{\mathrm{T}}$ are in "general positions" (Koenker, 2005, Section 2.2.2) and the true model contains at least one continuous covariate.

Assumption 3.3. *For every k and m, the conditional probability density of $\varepsilon_{km} = Y_k - X_k^{\mathrm{T}} \theta_{km}^*$ given X_k, say $f_{km}(\cdot|X_k)$, is uniformly bounded and bounded away from zero in a neighborhood around zero, and has a derivative $f'_{km}(\cdot|X_k)$ which is uniformly bounded in a neighborhood around zero.*

Assumption 3.4. *The true model size $q_n = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$.*

Assumption 3.5. *There exist positive constants c_2 and M_4 such that $2c_1 < c_2 \leq 1$ and*

$$n^{(1-c_2)/2} \min_{1 \leq j \leq q_n} \|\theta^{*(j)}\|_1 \geq M_4.$$

Assumptions 3.1 and 3.2 guarantee good behavior of the design matrices and existence of solutions to the quantile regression. Assumption 3.3 relates to the distributions of the random errors. These assumptions are weaker than requiring specific distributions. Assumption 3.4 regulates the

growth rate of the true model size. This is a standard assumption for linear models with a diverging number of parameters. Assumption 3.5 excludes situations where the nonzero parameters decay too fast. Conditions similar to Assumptions 3.1, 3.2, 3.3, 3.4 and 3.5 were required in Wang et al. (2012) for single experiments with single quantile.

Define the oracle estimator $\widehat{\theta}$ to be any local minimizer of $\ell_n(\theta)$ subject to $\|\widehat{\theta}^{(j)}\| = 0$ for $q_n < j \leq p_n$. The following theorem states model selection consistency. This means, with probability tending to one, that the oracle estimator can be obtained by minimizing the objective function $\Gamma_{\lambda_n}(\theta)$.

Theorem 3.1. *Let $S(\lambda_n)$ denote the set of local minimizers of $\Gamma_{\lambda_n}(\theta)$. Under Assumption 3.1, 3.2, 3.3, 3.4 and 3.5, we have $\mathrm{pr}\{\widehat{\theta} \in S(\lambda_n)\} \to 1$ as $n \to \infty$, if $\lambda_n = o(n^{-(1-c_2)/2})$, $n^{-1/2} q_n = o(\lambda_n)$ and $n^{-1} \log p_n = o(\lambda_n^2)$.*

The next theorem, Theorem 3.2, gives the asymptotic distribution of the nozero part of the oracle estimator from Theorem 3.1. We first introduce some notations: for every k, m and i, denote

$$
\begin{aligned}
&\varepsilon_{kmi} = Y_{ki} - X_{ki}^{\mathrm{T}} \theta_{km}^*,\ \varepsilon_{km} = (\varepsilon_{km1}, \ldots, \varepsilon_{kmn})^{\mathrm{T}},\ \varepsilon = (\varepsilon_{11}^{\mathrm{T}}, \ldots, \varepsilon_{1M}^{\mathrm{T}}, \ldots, \varepsilon_{K1}^{\mathrm{T}}, \ldots, \varepsilon_{KM}^{\mathrm{T}})^{\mathrm{T}},\\
&\psi_{kmi}(\varepsilon) = \tau_m - I(\varepsilon_{kmi} < 0),\ \psi_{nkm}(\varepsilon) = (\psi_{km1}(\varepsilon), \ldots, \psi_{kmn}(\varepsilon))^{\mathrm{T}},\\
&\psi_{nk}(\varepsilon) = (\psi_{nk1}(\varepsilon)^{\mathrm{T}}, \ldots, \psi_{nkM}(\varepsilon)^{\mathrm{T}})^{\mathrm{T}},\ \psi_n(\varepsilon) = (\psi_{n1}(\varepsilon)^{\mathrm{T}}, \ldots, \psi_{nK}(\varepsilon)^{\mathrm{T}})^{\mathrm{T}},\\
&H_n = E\{\psi_n(\varepsilon)\psi_n(\varepsilon)^{\mathrm{T}} | \mathcal{X}\} \text{ with } \mathcal{X} = \{X_{ki} : k = 1, \ldots, K, i = 1, \ldots, n\};\\
&B_{nkm} = \mathrm{diag}(f_{km}(0|X_{k1}), \ldots, f_{km}(0|X_{kn})),\ B_{nk} = \mathrm{diag}(B_{nk1}, \ldots, B_{nkM}),\\
&B_n = \mathrm{diag}(B_{n1}, \ldots, B_{nK});\ \widehat{\theta}_{kma} = (\widehat{\theta}_{km1}, \ldots, \widehat{\theta}_{kmq_n})^{\mathrm{T}},\\
&\widehat{\theta}_a = (\widehat{\theta}_{11a}^{\mathrm{T}}, \ldots, \widehat{\theta}_{1Ma}^{\mathrm{T}}, \ldots, \widehat{\theta}_{K1a}^{\mathrm{T}}, \ldots, \widehat{\theta}_{KMa}^{\mathrm{T}})^{\mathrm{T}},\ \theta_a^* = (\theta_{11a}^{*\mathrm{T}}, \ldots, \theta_{1Ma}^{*\mathrm{T}}, \ldots, \theta_{K1a}^{*\mathrm{T}}, \ldots, \theta_{KMa}^{*\mathrm{T}})^{\mathrm{T}}.
\end{aligned}
$$

We will show the asymptotic normality of the oracle estimator $\widehat{\theta}_a$.

Theorem 3.2. *Let $n^* = n \times M \times K$, $q_n^* = q_n \times M \times K$. Denote $X_a = diag(I_M \otimes X_{1\cdot a}, \ldots, I_M \otimes X_{K\cdot a})$ as a $n^* \times q_n^*$ block diagonal matrix, $R_n = n^{-1} X_a^{\mathrm{T}} B_n X_a$, $S_n = n^{-1} X_a^{\mathrm{T}} H_n X_a$ and $\Sigma_n =$*

$R_n^{-1} S_n R_n^{-1}$. *Consider a* $s \times q_n^*$ *matrix* A_n *with* s *fixed and* $A_n A_n^{\mathrm{T}} \to G$, *a positive definite matrix, then*

$$n^{1/2} A_n \Sigma_n^{-1/2} (\widehat{\theta}_a - \theta_a^*) \to N(0, G)$$

in distribution under Assumptions 3.1-3.4 and the condition that $\lambda_{\min}(S_n)$ *is uniformly bounded away from zero.*

Theorem 3.1 and 3.2 establish the model selection consistency and asymptotic normality of our estimator, even though the experiments are correlated. This makes it possible to aggregate information from multiple experiments, rather than ignoring the correlation and analyzing each experiment separately.

Remark 3.1. *A special case of (1.4) is homogeneous models, where the error* U_k *is independent of the covariate* X_k *for* $k = 1, \ldots, K$. *Then the conditional quantile* $Q_{\tau_m}(Y_k | X_k) = X_k^{\mathrm{T}} \alpha_k^* + b_{km}^*$, *where* b_{km}^* *is the* $\tau_m \times 100\%$ *quantile of* U_k, *has the same slope* α_k^* *through different quantile levels and the only difference is the intercepts* b_{km}^*. *Therefore, in each experiment, we can use multiple quantile levels to estimate this single slope, i.e., the penalized estimator* $(\widehat{\alpha}^{\mathrm{T}}, \widehat{b}^{\mathrm{T}})^{\mathrm{T}}$ *is obtained by minimizing*

$$\Pi_{\lambda_n}(\alpha, b) = n^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki}^{\mathrm{T}} \alpha_k - b_{km}) + \sum_{j=1}^{p_n} \Omega_{\lambda_n}(\|\alpha^{(j)}\|_1), \tag{3.3}$$

where $\alpha = (\alpha_1^{\mathrm{T}}, \ldots, \alpha_K^{\mathrm{T}})^{\mathrm{T}}$, $b = (b_{11}, \ldots, b_{1M}, \ldots, b_{K1}, \ldots, b_{KM})^{\mathrm{T}}$ *and* $\alpha^{(j)} = (\alpha_{1j}, \ldots, \alpha_{Kj})^{\mathrm{T}}$ *for* $j = 1, \ldots, p_n$. *For variable selection, we are interested in specifying the set* $\{j : 1 \le j \le p_n, \|\alpha^{*(j)}\| > 0\}$. *This is an extension of the composite quantile regression approach proposed by Zou & Yuan (2008a), to the situation with multiple experiments and high dimensional data. Selection consistency and asymptotic normality of estimators that minimize (3.3) can be established similarly to Theorem 3.1 and 3.2. In Section 3.3.2, we provide simulation results for homogeneous models. We will use a Bayesian information criterion, which is an analogue to the MQBIC from*

Section 3.2, to choose the tunining parameter λ_n in (3.3). See Section 3.3.2 for details.

3.2 Multiple quantile Bayesian information criterion

To select the correct model, we will use an information criterion that can balance the goodness-of-fit and the complexity of models. By applying this information criterion to competing models, the true model can be identified with probability approaching one. In the context of quantile regression, Lee et al. (2014) developed a Bayesian information criterion with a diverging number of predictors. Their method considers one single quantile and deals with data from one single experiment. We improve its power of model selection by aggregating information from additional (different) experiments with multiple quantiles. Since in our setting all the experiments share the same predictors, the aggregation makes it possible to identify predictors that have influence in any of the experiments.

Consider the multiple quantile Bayesian information criterion of a submodel $\mathcal{D} \subset \{1, 2, \ldots, p_n\}$,

$$\text{MQBIC}(\mathcal{D}) = \log\Big\{ \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{km\mathcal{D}}) \Big\} + (2n)^{-1} |\mathcal{D}| T_n \log n,$$

where $\widehat{\theta}_{km\mathcal{D}} = \arg\min_{\theta \in \mathbb{R}^{|\mathcal{D}|}} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}} \theta)$ for $k = 1, \ldots, K$ and $m = 1, \ldots, M$, $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$, T_n is a sequence of positive constants diverging to infinity as n increases. The notation $X_{ki\mathcal{D}}$ refers to the subvectors of $X_{ki\cdot}$, which only contain the components with subscripts in $\mathcal{D}$. We set an upper bound on the cardinality of competing models, say d_n, and search for the best model among submodels whose cardinality is smaller or equal to d_n. Define $\mathcal{D}^* = \{1, 2, \ldots, q_n\}$ as the subset of $\{1, \ldots, p_n\}$ corresponding to the true model, and $\mathcal{M} = \{\mathcal{D} \subset \{1, \ldots, p_n\} : |\mathcal{D}| \leq d_n\}$ as the set of all competing models. The first part of the MQBIC represents the goodness-of-fit, while the second term is a penalty on the model complexity. To obtain model selection consistency of MQBIC, we need the following assumptions, additionally to some of the assumptions from Section 3.1:

Assumption 3.6. *For every k, there are constants $0 < M_5 \leq M_6$ such that for any $\mathcal{D} \subset$*

$\{1, \ldots, p_n\}$ *the matrix* $X_{k\cdot\mathcal{D}} = (X_{k1\mathcal{D}}, \ldots, X_{kn\mathcal{D}})^{\mathrm{T}}$ *satisfies*

$$M_5 \leq \min_{|\mathcal{D}|\leq 2d_n} \lambda_{\min}(n^{-1}X_{k\cdot\mathcal{D}}^{\mathrm{T}}X_{k\cdot\mathcal{D}}) \leq \max_{|\mathcal{D}|\leq 2d_n} \lambda_{\max}(n^{-1}X_{k\cdot\mathcal{D}}^{\mathrm{T}}X_{k\cdot\mathcal{D}}) \leq M_6.$$

Assumption 3.7. *The full model size* p_n *is of order* $p_n = O(n^{c_3})$ *for some* $c_3 > 0$*; the true model size* q_n *is fixed,* $q_n = q$*, and satisfies* $q \leq d_n = O(n^{c_4})$ *for some* $0 < c_4 < 1/2$*.*

Assumption 3.8. *The sequence* T_n *satisfies* $T_n \to \infty$ *and* $n^{-1}T_n \log n \to 0$*.*

Assumption 3.9. *The summation of the check functions* $n^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(\varepsilon_{kmi})$ *is bounded away from zero and infinity with probability tending to one.*

Assumption 3.6 extends Assumption 3.2 for the true model to all candidate models. This is common for scenarios with more regression parameters than observations, i.e., $p_n > n$. In Assumption 3.7, the true model size is fixed for a technical difficulty in handling the maximum of $|\mathcal{D}\backslash\mathcal{D}^*|^{-1}|n^{-1}\sum_{i=1}^{n}\{\rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}}) - \rho_m(Y_{ki} - X_{ki\mathcal{D}^*}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}^*})\}|$ over the set of overfitted models $\{\mathcal{D} \in \mathcal{M} : \mathcal{D}^* \subset \mathcal{D}, \mathcal{D} \neq \mathcal{D}^*\}$ (Lee et al., 2014). Assumption 3.8 regulates the convergence rate of the sequence T_n. Assumption 3.9 is made for convenience in the proof.

In the following theorem we show that the true model has, with probability tending to one, the smallest MQBIC value among all candidates.

Theorem 3.3. *If Assumptions 3.1, 3.3 and 3.6-3.9 hold, then with probability tending to one, the true model can be selected by minimizing the MQBIC, that is*

$$\lim_{n\to\infty} \mathrm{pr}\Big\{\min_{\mathcal{D}\in(\mathcal{M}\backslash\{\mathcal{D}^*\})} \textit{MQBIC}(\mathcal{D}) > \textit{MQBIC}(\mathcal{D}^*)\Big\} = 1.$$

Theorem 3.3 establishes model selection consistency of the MQBIC for data from multiple dependent sources, which provides another approach to identify the true underlying model. Compared with the result in Lee et al. (2014), aggregating information from multiple data sources and quantiles helps to better detect the predictors' effect on responses in the context of data integration. In the MQBIC approach estimation and model selection are separate processes . This is different

from minimizing the objective function in Section 3.1, which is a one-step procedure. Moreover, the MQBIC is also a useful tool to select the tuning parameter λ_n for the penalized estimation process in Section 3.1. The details are shown in Section 3.3.

3.3 Simulations

3.3.1 Basic Settings

In this section we study the numerical performance of our estimators in two homogeneous settings with independent errors and covariates. The objective function for this scenario is provided in (3.3) in Remark 3.1. We also consider two heterogeneous settings. Here the error terms depend on the covariates and estimators are obtained by minimizing (3.2). We study two different grouping structures, namely complete and incomplete grouping. Complete grouping means that parameters of the same predictor can only be either all zero or nonzero, while in the latter case, a group may contain both zero and nonzero predictors.

In each of the four cases, the number of experiments is $K = 2$, the sample size is $n = 300$ and the number of predictors is $p = 400$ or 600. The nonzero parameters are drawn from a uniform distribution on $[0.05, 1]$ independently. For $K = 1, 2$ and $i = 1, \ldots, 300$, we generate random vectors X'_{ki} independently from a p-dimensional multivariate normal distribution with a mean zero and variance-covariance matrix Σ. The matrix Σ is a block diagonal matrix, whose diagonal entries are 10×10 matrices with (i, j) component being $0.8^{|i-j|}$. The X'_{ki}'s will be transformed to predictors X_{ki} in different ways for different situations. For $i = 1, \ldots, 300$, error terms (ξ_{1i}, ξ_{2i}) are drawn independently from a bivariate normal distribution with a mean zero and variance-covariance matrix Σ' with entries $\Sigma'_{11} = \Sigma'_{22} = 1$ and $\Sigma'_{12} = \Sigma'_{21} = 0.7$. Minimizing the objective functions uses the QICD algorithm proposed by Peng & Wang (2015).

In each situation we record three indices:

1. positive selection rate (PSR): the proportion of predictors affecting any quantiles of any responses that are selected, i.e., $|\widehat{\mathcal{A}} \cap \mathcal{A}|/|\mathcal{A}|$, where $\mathcal{A} = \{j : 1 \leq j \leq p, \|\theta^{*(j)}\| > 0\}$ and $\widehat{\mathcal{A}} = \{j : 1 \leq j \leq p, \|\widehat{\theta}^{(j)}\| > 0\}$ for the heterogeneous cases, $\mathcal{A} = \{j : 1 \leq j \leq$

$p, \|\alpha^{*(j)}\| > 0\}$ and $\widehat{\mathcal{A}} = \{j : 1 \le j \le p, \|\widehat{\alpha}^{(j)}\| > 0\}$ for the homogeneous cases, and for a set, $|\cdot|$ means its cardinality;

2. false discovery rate (FDR): the proportion of predictors affecting no response that are selected, i.e., $|\widehat{\mathcal{A}} \cap \mathcal{A}^c|/|\mathcal{A}^c|$;

3. absolute error (AE): the absolute estimating error defined by $K^{-1}\sum_{k=1}^{K}\sum_{j=1}^{p}|\widehat{\alpha}_{kj} - \alpha^*_{kj}|$ and $(KM)^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{j=1}^{p}|\widehat{\theta}_{kmj} - \theta^*_{kmj}|$ for homogeneous and heterogeneous models, respectively.

The data integration (DI) method is compared with the combined analysis with τ quantile (CA-τ). This method considers only one quantile τ; it analyzes the data from the two experiments separately and then combines the results. We present their averages over 100 simulated data sets. The standard deviations are also recorded; see the parentheses after the averages.

3.3.2 Homogeneous models

For homogeneous scenarios $M = 3$ quantiles are used in the objective function (3.3), $\tau_1 = 0.3$, $\tau_2 = 0.5$ and $\tau_3 = 0.7$. We set the predictors $X_{ki} = X'_{ki}$ and the responses $Y_{ki} = X_{ki}^{\mathrm{T}}\alpha^*_k + \xi_{ki}$ for $k = 1, 2$ and $i = 1, \ldots, 300$. We choose the tuning parameter λ from a fine grid Λ. For any $\lambda \in \Lambda$, let $\widehat{\alpha}_{\lambda,k} = (\widehat{\alpha}_{\lambda,k1}, \ldots, \widehat{\alpha}_{\lambda,kp})^{\mathrm{T}}$ and $\widehat{b}_{\lambda,km}$ denote the estimators obtained from minimizing the objective function (3.3) with the tuning parameter $\lambda_n = \lambda$, where $k = 1, 2$ and $m = 1, 2, 3$. Then let $\mathcal{D}_\lambda = \{j : 1 \le j \le p, \sum_{k=1}^{K}|\widehat{\alpha}_{\lambda,kj}| > 0\}$. We use

$$\widehat{\lambda} = \underset{\lambda\in\Lambda}{\arg\min}\Big[\log\Big\{\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{\lambda,ki}^{\mathrm{T}}\widehat{\alpha}_{\lambda,k} - \widehat{b}_{\lambda,km})\Big\} + (2n)^{-1}|\mathcal{D}_\lambda|(\log n)T\Big],$$

with $T = \log p$ to obtain the final estimators. This criterion is an adaption of (2.10) in Lee et al. (2014) to multiple experiments; $T = \log p$ is recommended by the article.

Table 3.2 gives the simulation results of a model with a completely grouping structure. The nonzero parameters are α^*_{11}, α^*_{16}, $\alpha^*_{1(12)}$, $\alpha^*_{1(15)}$, $\alpha^*_{1(20)}$ and α^*_{21}, α^*_{26}, $\alpha^*_{2(12)}$, $\alpha^*_{2(15)}$, $\alpha^*_{2(20)}$. The data integration (DI) method is compared with the combined analysis (CA). The DI achieves higher pos-

itive selection rates by aggregating information from the two data sets. Moreover, the DI method significantly reduces absolute estimating errors by using multiple quantiles.

In Table 3.3 we present the simulation results for the scenario from Table 3.2, but with an incompletely grouping structure. The nonzero parameters are α^*_{11}, α^*_{13}, α^*_{16}, α^*_{19}, $\alpha^*_{1(12)}$, $\alpha^*_{1(20)}$ and α^*_{21}, α^*_{23}, $\alpha^*_{2(12)}$, $\alpha^*_{2(15)}$, $\alpha^*_{2(20)}$, $\alpha^*_{2(40)}$. The DI method again outperforms the CA: it achievies higher positive selection rates and lower absolute errors. This confirms that it selects important predictors more effectively and provides more preciese estimates than competing approaches.

Table 3.2: Simulated positive selection rates, false discovery rates and absolute errors of the data integration and the combined analysis for homogeneous models with the complete grouping structure

	$p = 400$			$p = 600$		
	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	99.2 (3.9)	0.1 (0.2)	0.4 (0.1)	99.4 (3.4)	0.1 (0.2)	0.4 (0.2)
CA-0.3	97.0 (7.1)	0.1 (0.1)	0.8 (0.2)	95.2 (8.6)	0.1 (0.1)	0.8 (0.2)
CA-0.5	98.0 (6.0)	0.1 (0.1)	0.7 (0.2)	98.8 (4.8)	0.1 (0.1)	0.7 (0.3)
CA-0.7	96.4 (7.7)	0.1 (0.1)	0.8 (0.3)	95.6 (8.3)	0.1 (0.1)	0.8 (0.3)

Here p, the full model size; DI, data integration method; CA-τ, combined analysis with τ quantile; PSR, positive selection rate; FDR, false discovery rate; AE, absolute error $K^{-1}\sum_{k=1}^{K}\sum_{j=1}^{p}|\widehat{\alpha}_{kj} - \alpha^*_{kj}|$.

3.3.3 Heterogeneous models

We now test the performance of our estimator in the heterogeneous scenario. We use $M = 5$ quantiles in the objective function (3.2), $\tau_1 = 1/6, \tau_2 = 2/6, \ldots, \tau_5 = 5/6$. We choose the tuning parameter λ from a grid Λ. For any $\lambda \in \Lambda$, let $\widehat{\theta}_{\lambda,km} = (\widehat{\theta}_{\lambda,km1}, \ldots, \widehat{\theta}_{\lambda,kmp})^{\mathrm{T}}$ denote the estimators obtained from minimizing the objective function (3.2) with $\lambda_n = \lambda$, where $k = 1, 2$ and $m = 1, 2, 3, 4, 5$. Further let $\mathcal{D}_\lambda = \{j : 1 \le j \le p, \sum_{k=1}^{K}\sum_{m=1}^{M}|\widehat{\theta}_{\lambda,kmj}| > 0\}$. The minimizer of

Table 3.3: Simulated positive selection rates, false discovery rates and absolute errors of the data integration and the combined analysis for homogeneous models with the incomplete grouping structure

	$p = 400$			$p = 600$		
	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	96.4 (6.5)	0 (0.1)	0.8 (0.3)	95.3 (6.1)	0 (0)	0.8 (0.3)
CA-0.3	90.8 (6.6)	0 (0.1)	1.3 (0.4)	90.3(6.5)	0 (0.1)	1.3 (0.4)
CA-0.5	92.3 (6.6)	0 (0.1)	1.2 (0.3)	91.3 (6.0)	0 (0.1)	1.2 (0.3)
CA-0.7	91.0 (7.1)	0 (0.1)	1.3 (0.5)	90.3 (6.8)	0 (0.1)	1.3 (0.5)

Here p, the full model size; DI, data integration method; CA-τ, combined analysis with τ quantile; PSR, positive selection rate; FDR, false discovery rate; AE, absolute error $K^{-1}\sum_{k=1}^{K}\sum_{j=1}^{p}|\widehat{\alpha}_{kj}-\alpha^*_{kj}|$.

the MQBIC, i.e.,

$$\widehat{\lambda} = \arg\min_{\lambda\in\Lambda}\Big[\log\Big\{\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki}-X^{\mathrm{T}}_{\lambda,ki}\widehat{\theta}_{\lambda,km})\Big\}+(2n)^{-1}|\mathcal{D}_\lambda|(\log n)T\Big], \tag{3.4}$$

is used to obtain the final estimators. Here we set $T = \log p$ or $(\log p)/6$ and examine how this affects the performance of the method.

Table 3.4 gives the simulation results of a model with a completely grouping structure. The nonzero parameters are α^*_{11}, α^*_{16}, $\alpha^*_{1(12)}$, $\alpha^*_{1(15)}$, $\alpha^*_{1(20)}$ and α^*_{21}, α^*_{26}, $\alpha^*_{2(12)}$, $\alpha^*_{2(15)}$, $\alpha^*_{2(20)}$. Let $\Phi(\cdot)$ be the distribution function of a standard normal variable. For $k = 1, 2$ and $i = 1, \ldots, 300$ the predictors are $X_{ki3} = \Phi(X'_{ki3})$ and $X_{kij} = X'_{kij}$ for $j \neq 3$. The responses are $Y_{ki} = X^{\mathrm{T}}_{ki}\alpha^*_k + 0.7\xi_{ki}X_{ki3}$. Among all the methods, the DI gives the highest positive selection rates and the lowest false discovery rates. Moreover, it has the lowest absolute errors. We also observe that the DI method is not significantly affected by the different values of T.

In Table 3.5 we present the simulation results of a model with ian ncompletely grouping structure. The nonzero parameters are α^*_{14}, α^*_{16}, α^*_{19}, $\alpha^*_{1(12)}$, $\alpha^*_{1(15)}$, $\alpha^*_{1(20)}$ and α^*_{21}, α^*_{26}, $\alpha^*_{2(12)}$, $\alpha^*_{2(15)}$, $\alpha^*_{2(20)}$, $\alpha^*_{2(25)}$. For $i = 1, \ldots, 300$, the predictors of the first experiment are $X_{1i1} = \Phi(X'_{1i1})$

and $X_{1ij} = X'_{1ij}$ for $j \neq 1$. The predictors of the second experiment are $X_{2i3} = \Phi(X'_{2i3})$ and $X_{2ij} = X'_{2ij}$ for $j \neq 3$. The responses are generated by $Y_{1i} = X_{1i}^{\mathrm{T}}\alpha_1^* + 0.7\xi_{1i}X_{1i1}$ and $Y_{2i} = X_{2i}^{\mathrm{T}}\alpha_2^* + 0.7\xi_{2i}X_{2i3}$. Here the DI still has higher positive selection rates and lower ralse discovery rates, and produces smaller absolute errors than its competitors. As in the previous table, the choice of T is only of marginal importance for the performance of the DI method.

Table 3.4: Simulated positive selection rates, false discovery rates and absolute errors of the data integration and the combined analysis for heterogeneous models with the complete grouping structure

		$p = 400$			$p = 600$		
		PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
(a)	DI	98.3 (5.0)	0 (0)	0.1 (0)	98.2 (5.2)	0 (0)	0.1 (0.1)
	CA-(2/6)	82.8 (10.4)	0.1 (0.1)	0.5 (0.1)	80.0(11.4)	0 (0.1)	0.6 (0.1)
	CA-(3/6)	80.0 (6.7)	0 (0.1)	0.2 (0.1)	79.3 (7.2)	0 (0)	0.2 (0.1)
	CA-(4/6)	83.2 (11.7)	0.1 (0.1)	0.5 (0.1)	80.2 (10.0)	0 (0.1)	0.6 (0.1)
(b)	DI	99.8 (1.7)	0 (0.1)	0.1 (0)	99.8 (1.7)	0 (0)	0.1 (0)
	CA-(2/6)	96.1 (7.0)	1.9 (1.3)	0.3 (0.1)	94.3 (7.9)	1.0 (0.8)	0.4 (0.1)
	CA-(3/6)	83.2 (1.7)	0.6 (0.5)	0.2 (0)	83.3 (0)	0.4 (0.3)	0.2 (0)
	CA-(4/6)	97.5 (6.0)	1.7 (1.1)	0.3 (0.1)	94.0 (8.0)	1.0 (0.8)	0.4 (0.1)

Here p, the full model size; DI, data integration method; CA-τ, combined analysis with τ quantile; PSR, positive selection rate; FDR, false discovery rate; AE, absolute error $(KM)^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{j=1}^{p}|\widehat{\theta}_{kmj} - \theta^*_{kmj}|$. The parameter T in the criterion (3.4) is (a) $\log p$ or (b) $(\log p)/6$.

3.4 Real data analysis

This section analyzes data sets of financial market indices from the `R` package `FusionLearn`, which contain three correlated indices: the VIX index, the S&P 500 index and the Dow Jones index. The VIX and the S&P 500 are negatively correlated, while the S&P 500 and the Dow Jones

Table 3.5: Simulated positive selection rates, false discovery rates and absolute errors of the data integration and the combined analysis for heterogeneous models with the incomplete grouping structure

		$p = 400$			$p = 600$		
		PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
(a)	DI	92.7 (5.3)	0 (0.1)	0.2 (0.1)	93.9 (5.8)	0 (0)	0.2 (0.1)
	CA-(2/6)	83.3 (5.6)	0.1 (0.1)	0.6 (0.1)	84.1(6.0)	0 (0.1)	0.6 (0.1)
	CA-(3/6)	88.9 (0)	0 (0.1)	0.3 (0.1)	83.9 (5.6)	0 (0.1)	0.3 (0.1)
	CA-(4/6)	85.0 (7.3)	0.1 (0.2)	0.5 (0.1)	84.7 (6.1)	0 (0.1)	0.6 (0.1)
(b)	DI	98.8 (3.5)	0.1 (0.1)	0.2 (0.1)	98.3 (4.0)	0 (0.1)	0.2 (0.1)
	CA-(2/6)	94.9 (5.6)	2.1 (1.2)	0.4 (0.1)	92.3 (5.6)	1.1 (0.7)	0.4 (0.1)
	CA-(3/6)	88.7 (1.6)	0.7 (0.5)	0.2 (0)	88.7 (1.6)	0.4 (0.3)	0.2 (0)
	CA-(4/6)	95.7 (5.7)	2.0 (1.4)	0.3 (0.1)	93.6 (5.5)	1.3 (0.9)	0.4 (0.1)

Here p, the full model size; DI, data integration method; CA-τ, combined analysis with τ quantile; PSR, positive selection rate; FDR, false discovery rate; AE, absolute error $(KM)^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{j=1}^{p}|\widehat{\theta}_{kmj} - \theta^*_{kmj}|$. The parameter T in the criterion (3.4) is (a) $\log p$ or (b) $(\log p)/6$.

are positively correlated (Gao & Carroll, 2017). The covariates are 46 major international equity indices, North American bond indices and major commodities indices. In the analysis, $\log(V_t / V_y) \times 100$ of each index is used, where V_t is today's value and V_y means yesterday's value. In the training data set, there are 232 records of three years market performances with three-day spacing between the values. As indicated in Gao & Carroll (2017), the values are not autocorrelated at a 5% significance level.

We fit a heterogeneous model and minimize the objective function (3.2) to select covariates and estimate parameters. Quantiles $\tau_m = m/20$ for $m = 1, 2, \ldots, 19$ and two different penalties, the SCAD and minimax concave penalty (MCP), are used. The tuning parameter of the penalty is chosen by (3.4) with $T = \log p$. The SCAD selects 4 covariates, which are the same as the 4 covariates selected by the MCP. For comparison, we also consider the combined analysis (CA),

which deals with the multiple quantiles and data sets seperately and then combines the results, and the full model including all the 46 covariates. The five fitted models are employed to make predictions on a validation data set of another 464 records. Prediction errors

$$\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki}-X_{ki}^{\mathrm{T}}\widehat{\theta}_{km}-\widehat{b}_{km}) \tag{3.5}$$

for $k = 1, 2, 3$ are displayed in Table 3.6, where $\widehat{b}_{km}$ is the estimated intercept in $Q_{\tau_m}(Y_k|X_k)$. The DI method with either the SCAD or MCP outperforms the other approaches while the SCAD and MCP give close prediction errors. Moreover, the DI method achieves more sparsity since it generates models with smaller sizes than those from the CA. This real data example again clearly demonstrates the advantage of our method .

Table 3.6: Prediction errors and model sizes of selected subset models and the full model

	PE			MS
	VIX	S&P 500	Dow Jones	
DI with SCAD	10045.8	524.9	306.9	4
DI with MCP	10026.5	522.7	308.8	4
CA with SCAD	10139.9	637.6	398.6	23
CA with MCP	10115.8	637.8	391.0	19
Full model	13408.5	644.0	663.4	46

Here PE, prediction error (3.5); MS, model size; DI, data integration method; CA, combined analysis; SCAD, smoothly clipped absolute deviation; MCP, minimax concave penalty.

4. CONCLUSIONS AND DISCUSSIONS

4.1 Conclusions and discussions of Chapter 2

Chapter 2 has developed an efficient estimator for $E\{h(X,Y)\}$, the expectation of a known square-integrable function of the response and covariate, in parametric regression models with the response missing at random. This describes the entire relation between the response and covariate. The efficiency of the estimator has been proved by showing the expansion of our estimator matches the canonical gradient of the functional $E\{h(X,Y)\}$. Simulations and a real data example verify the optimality of our estimator by comparing it with other competing methods.

Our estimator uses the single imputation technique. Since $h(X_i,Y_i)$ cannot be used to estimate $E\{h(X,Y)\}$ if the response Y_i is missing, we use instead estimators of the conditional expectation $E\{h(X,Y)|X\}$ that do not involve the missing Y_i's (and also not surrogates for the missing Y_i's). As pointed out by a referee, although our approach is *asymptotically* optimal (efficient), when the sample size is small, our method may benefit from applying multiple imputation, which uses a reasonable imputation model to generate several complete data sets, analyzes these sets separately and then pools the results. Therefore exploring how to improve the performance of the estimator by multiple imputation is an interesting question, which is left for further study.

Furthermore, robustness to model misspecification is also a concern. Because we assume a parametric regression function $r_\vartheta(X)$ and exploit its structure, performance of our approach depends on the correct specification of the model. Hu et al. (2010) developed an estimator for $E(Y)$ using a parametric model for $\pi(X)$, which is robust to misspecification of the regression fucntion. However, the robustness relies on correct specification of $\pi(X)$, and only if both $r_\vartheta(X)$ and $\pi(X)$ are correctly specified will the estimator be efficient. These are strong requirements. In this regard, strategies to deal with misspecification under mild conditions are worth deeper investigations.

4.2 Conclusions and discussions of Chapter 3

In Chapter 3, we have introduced a quantile regression approach to the scenario of data integration with high dimensional data. We have proposed a penalized estimator and an information criterion, which aggregate information from multiple experiments, to select variables and to estimate model parameters. The asymptotic properties of these approaches have been proved. Our method successfully takes grouping structures across experiments into account. It provides a global picture of the relationship between predictors and responses by considering multiple quantile levels simultaneously.

In practice quality and importance of data may vary from one source to another. Therefore, a weighted version of the loss function (3.1), i.e.,

$$\ell_n^{(w)}(\theta) = n^{-1} \sum_{k=1}^{K} w_k \sum_{m=1}^{M} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki}^{\mathrm{T}} \theta_{km})$$

with a weight vector $w = (w_1, \ldots, w_k)^{\mathrm{T}}$, may improve our estimator, which uses uniform weights. It will be interesting to explore how to specify proper weights for data from different experiments and estimate these weights.

The nonconvex penalty function associated with the L_1-norm has different properties than that associated with the L_2-norm. Jiang & Huang (2015) proved that the penalty associated with the L_1-norm has the bi-level selection property (in and between groups) when the least square approach is used. In the simulations of Section 3.3 we have seen that the nonconvex penalty with the L_1-norm performs well when zero and nonzero parameters exist in the same group. Its theoretical properties in the quantile regression setting, however, still need to be investigated in greater detail.

REFERENCES

BATES, D. M. & WATTS, D. G. (1998). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

BELLONI, A. & CHERNOZHUKOV, V. (2011). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39**, 82–130.

CHENG, P. E. (1990). Applications of kernel regression estimation: survey. *Communications in statistics-theory and methods* **19**, 4103–4134.

CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* **89**, 81–87.

DEVROYE, L. P. & WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *The Annals of Statistics* **8**, 231–239.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

GAO, X. & CARROLL, R. J. (2017). Data integration with high dimensionality. *Biometrika* **104**, 251–272.

HAMMER, S. M., KATZENSTEIN, D. A., HUGHES, M. D., GUNDACKER, H., SCHOOLEY, R. T., HAUBRICH, R. H., HENRY, W. K., LEDERMAN, M. M., PHAIR, J. P., NIU, M., HIRSCH, M. S. & MERIGAN, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335**, 1081–1090.

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24**, 726–748.

HIRANO, K., IMBENS, G. W. & RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

HU, Z., FOLLMANN, D. A. & QIN, J. (2010). Semiparametric dimension reduction estimation for mean response with missing data. *Biometrika* **97**, 305–319.

JIANG, D. & HUANG, J. (2015). Concave 1-norm group selection. *Biostatistics* **16**, 252–267.

KNIGHT, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Annals of Statistics* , 755–770.

KOENKER, R. (2005). *Quantile regression*. Cambridge, UK: Cambridge University Press.

KOENKER, R. & BASSETT, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society* , 33–50.

LEE, E. R., NOH, H. & PARK, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association* **109**, 216–229.

LITTLE, R. J. A. & RUBIN, D. B. (2019). *Statistical analysis with missing data*, vol. 793. New York: Wiley.

MATLOFF, N. S. (1981). Use of regression functions for improved estimation of means. *Biometrika* **68**, 685–689.

MÜLLER, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *The Annals of Statistics* **37**, 2245–2277.

MÜLLER, U. U. & SCHICK, A. (2017). Efficiency transfer for regression models with responses missing at random. *Bernoulli* **23**, 2693–2719.

MÜLLER, U. U., SCHICK, A. & WOLFGANG, W. (2006). Imputing responses that are not missing. In *Probability, Statistics and Modelling in Public Health*, H. C. Nikulin M., Commenges D., ed. Symposium in Honor of Marvin Zelen, Springer.

MÜLLER, U. U. & VAN KEILEGOM, I. (2012). Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics* **6**, 1200–1219.

NADARAYA, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications* **9**, 141–142.

PENG, B. & WANG, L. (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics* **24**, 676–694.

SCHICK, A. (1993). On efficient estimation in regression models. *The Annals of Statistics* **21**,

1486–1521.

SEBER, G. A. F. & WILD, C. J. (1989). *Nonlinear Regression*. New York: J. Wiley & Sons.

SHERWOOD, B. & WANG, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics* **44**, 288–317.

SIMONOFF, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

TANG, M.-L., TANG, N.-S., ZHAO, P.-Y. & ZHU, H. (2018). Efficient robust estimation for linear models with missing response at random. *Scandinavian Journal of Statistics* **45**, 366–381.

WAND, M. P. & SCHUCANY, W. R. (1990). Gaussian-based kernels. *Canadian Journal of Statistics* **18**, 197–204.

WANG, L., WU, Y. & LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214–222.

WANG, Q. & RAO, J. N. K. (2001). Empirical likelihood for linear regression models under imputation for missing responses. *Canadian Journal of Statistics* **29**, 597–608.

WANG, Q. & RAO, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics* **30**, 896–924.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* , 359–372.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical association* **101**, 1418–1429.

ZOU, H. & YUAN, M. (2008a). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36**, 1108–1126.

ZOU, H. & YUAN, M. (2008b). Regularized simultaneous model selection in multiple quantiles

regression. *Computational Statistics & Data Analysis* **52**, 5296–5304.

APPENDIX A

TECHNICAL DETAILS OF CHAPTER 2

A.1 An auxiliary lemma

Lemma A.1. *Let $(X_1, V_1), \ldots, (X_n, V_n)$ be i.i.d. copies of a base observation (X, V), where X satisfies Assumption 2.3 and V is a q-dimensional random vector. For some function $g(x, v) : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$, set $m(x) = E\{g(X, V)|X = x\}$. Suppose further that the distribution of (X, V) has a joint density and that Assumptions 2.4, 2.5 and 2.6 are satisfied.*

1. *If $m(x)$ is $d + 1$ times continuously differentiable over $\mathcal{I}$, then*

$$\sup_{x\in\mathcal{I}} \left|E\{g(X, V)K_b(X, x)\} - f(x)m(x)\right| = o_p(n^{-1/2}).$$

2. *Further, if $E\{g^2(X, V)\}$ is finite, then*

$$\sup_{x\in\mathcal{I}} \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i, V_i)K_b(X_i, x) - f(x)m(x)\right| = o_p(n^{-1/4}).$$

Proof of Lemma A.1: Let f_2 denote the joint density of (X, V) and $f(\cdot|x)$ denote the conditional density of V given $X = x$. For the proof of the first part we write $\mu(x) = E\{g(X, V)K_b(X, x)\}$ and, using substitution, obtain

$$\begin{aligned}
\mu(x) &= \int_{\mathcal{I}}\int_{\mathbb{R}^q} g(u, v)b^{-d}K\Big(b^{-1}(u - x), x\Big)f_2(u, v)dvdu \\
&= \int_{\mathcal{S}_b(x)}\int_{\mathbb{R}^q} g(bs + x, v)K(s, x)f_2(bs + x, v)dvds \\
&= \int_{\mathcal{S}_b(x)} K(s, x)f(bs + x)\int_{\mathbb{R}^q} g(bs + x, v)f(v|bs + x)dvds \\
&= \int_{\mathcal{S}_b(x)} K(s, x)f(bs + x)m(bs + x)ds.
\end{aligned}$$

A Taylor expansion gives that for $s = (s_1, \ldots, s_d)^{\mathrm{T}} \in \mathcal{S}_b(x)$,

$$f(bs+x)m(bs+x) = \sum_{|\alpha| \le d} \frac{D^\alpha\{f(x)m(x)\}}{\alpha!}(bs)^\alpha + \sum_{|\alpha|=d+1} R_\alpha(x,s)(bs)^\alpha$$

where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, $|\alpha| = \sum_{i=1}^d \alpha_i$, $\alpha! = \prod_{i=1}^d \alpha_i!$, $(bs)^\alpha = b^{|\alpha|}s^\alpha = b^{|\alpha|}\prod_{i=1}^d s_i^{\alpha_i}$ and

$$D^\alpha\{f(x)m(x)\} = \frac{\partial^{|\alpha|}\{f(x)m(x)\}}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}.$$

For example, if $x = (x_1, x_2)^{\mathrm{T}}$ and $s = (s_1, s_2)^{\mathrm{T}}$ are two-dimensional vectors, we have

$$\begin{aligned}&\sum_{|\alpha|=2} \frac{D^\alpha\{f(x)m(x)\}}{\alpha!}(bs)^\alpha \\ &= \frac{\partial^2 f(x)m(x)}{\partial x_1^2}\frac{(bs_1)^2}{2} + \frac{\partial^2 f(x)m(x)}{\partial x_2^2}\frac{(bs_2)^2}{2} + \frac{\partial^2 f(x)m(x)}{\partial x_1 \partial x_2}b^2 s_1 s_2.\end{aligned}$$

Since $f(x)$ and $m(x)$ are $d+1$ times continuously differentiable over $\mathcal{I}$, the term in the remainder is

$$R_\alpha(x,s) = \frac{|\alpha|}{\alpha!}\int_0^1 (1-t)^{|\alpha|-1} D^\alpha\{f(x+tbs)m(x+tbs)\}dt.$$

When $|\alpha| = d+1$, it follows that

$$\left|R_\alpha(x,s)\right| \le \frac{1}{\alpha!}\sup_{|\beta|=d+1}\sup_{w\in\mathcal{I}}\left|D^\beta\{f(w)m(w)\}\right| \le c \tag{A.1}$$

with $\beta \in \mathbb{R}^d$, because $f(x)$ and $m(x)$ are $d+1$ times continuously differentiable over $\mathcal{I}$. By Assumption 2.4 (ii) we have

$$\mu(x) = f(x)m(x) + b^{d+1}\int_{\mathcal{S}_b(x)} K(s,x)\sum_{|\alpha|=d+1} R_\alpha(x,s)s^\alpha ds,$$

which implies that

$$\begin{aligned}
&\sup_{x\in\mathcal{I}} |\mu(x) - f(x)m(x)| \\
&\leq \sup_{x\in\mathcal{I}} \Big\{ b^{d+1} \int_{\mathcal{S}_b(x)} \big|K(s,x)\big| \sum_{|\alpha|=d+1} \Big(\big|R_\alpha(x,s)\big| |s^\alpha| \Big) ds \Big\} \\
&\leq cb^{d+1} \sup_{x\in\mathcal{I}} \Big\{ \sum_{|\alpha|=d+1} \int \big|K(s,x)s^\alpha\big| ds \Big\} = O_p(b^{d+1}),
\end{aligned}$$

where the second step is because of (A.1), and the last step comes from Assumption 2.4 (i). Therefore, by Assumption 2.5, we have

$$\sup_{x\in\mathcal{I}} |\mu(x) - f(x)m(x)| = o_p(n^{-1/2}). \tag{A.2}$$

We now prove part (2). Analogously as in the derivation of Theorem 2 in Hansen (2008), where Y in the original proof is replaced by $g(X,V)$, we have

$$\sup_{x\in\mathcal{I}} \Big| \frac{1}{n} \sum_{i=1}^{n} g(X_i,V_i) K_b(X_i,x) - \mu(x) \Big| = O_p\Big(\Big(\frac{\log n}{nb^d} \Big)^{1/2} \Big) = o_p(n^{-1/4}). \tag{A.3}$$

The assumptions of that theorem are satisfied:

1. Assumptions 1 and 3 in Hansen (2008) hold true by Assumption 2.4 (i) and 2.4 (iii), respectively;

2. we have independent observations, so conditions (2), (4), (7) and (10) in Hansen (2008) are not needed, and (11) in that article simplifies to $\theta = 1$;

3. condition (5) in Hansen(2008) is satisfied by Assumption 2.3, and inspecting the proofs of Theorems 1 and 2 in Hansen (2008) reveals that condition (3) and (6) in that article can be replaced by the assumption that $g(X,Y)$ is square integrable for independent data;

4. equation (12) in Hansen (2008) is met by Assumption 2.5; equation (13) in that article is satisfied since the support $\mathcal{I}$ in (A.3) is compact.

Combining (A.2) and (A.3) gives the desired statement,

$$\sup_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i,V_i)K_b(X_i,x) - f(x)m(x)\Big| = o_p(n^{-1/4}).$$

□

A.2 Proof of Theorem 2.1

To prove this theorem we write $\widehat{H}_{np} = A + B_1$, where

$$A = \frac{1}{n}\sum_{i=1}^{n}[\chi(X_i) + Z_i\{h(X_i,Y_i) - \chi(X_i)\}], \tag{A.4}$$
$$B_1 = \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\{\widehat{\chi}(X_i) - \chi(X_i)\}.$$

We will show that B_1 and the term B_3 given below are asymptotically equivalent. This will be established using Assumptions 2.3, 2.4, 2.5 and 2.6 and Lemma A.1. Then we show that the leading term of $\widehat{H}_{np}$ stated in the theorem is an approximation of $A + B_3$.

We first introduce

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\{\phi(X_i) - \widetilde{\phi}(X_i)\}$$

with

$$\phi(X_i) = \frac{n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j,X_i)h(X_j,Y_j)}{\pi(X_i)f(X_i)},$$
$$\widetilde{\phi}(X_i) = \frac{n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j,X_i)\chi(X_i)}{\pi(X_i)f(X_i)}.$$

We can write B_1 in the form

$$B_1 = \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\frac{\sum_{j=1}^{n} Z_j K_b(X_j,X_i)\{h(X_j,Y_j) - \chi(X_i)\}}{\sum_{j=1}^{n} Z_j K_b(X_j,X_i)}.$$

By the second conclusion of Lemma A.1 we have

$$J_1 = \sup_{x\in\mathcal{I}} \Big|\frac{1}{n}\sum_{j=1}^{n} Z_j K_b(X_j, x)\{h(X_j, Y_j) - \chi(x)\}\Big| = o_p(n^{-1/4}),$$
$$J_2 = \sup_{x\in\mathcal{I}} \Big|\frac{1}{n}\sum_{j=1}^{n} Z_j K_b(X_j, x) - \pi(x)f(x)\Big| = o_p(n^{-1/4}),$$

and therefore

$$\begin{aligned}
|B_2 - B_1| &\leq \frac{1}{n}\sum_{i=1}^{n}\Big[(1-Z_i)\frac{|n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_i)\{h(X_j, Y_j) - \chi(X_i)\}|}{\pi(X_i)f(X_i)|n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_i)|} \\
&\quad \times\Big|\frac{1}{n}\sum_{j=1}^{n} Z_j K_b(X_j, X_i) - \pi(X_i)f(X_i)\Big|\Big] \\
&\leq J_1 J_2 \frac{1}{n}\sum_{i=1}^{n}\frac{1-Z_i}{\pi(X_i)f(X_i)|n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_i)|} \\
&\leq o_p(n^{-1/2})\Big\{\inf_{x\in\mathcal{I}}\pi(x)f(x)\Big\}^{-1}\Big\{\inf_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{j=1}^{n} Z_j K_b(X_j, x)\Big|\Big\}^{-1}.
\end{aligned}$$

where the two items are both bounded away from zero by Assumption 2.3 and (2.2). This shows

$$|B_2 - B_1| = o_p(n^{-1/2}). \tag{A.5}$$

Then we define B_3 as the conditional expectation of B_2 given the completely observed cases "$\mathcal{B}$", i.e.,

$$B_3 = E(B_2|\mathcal{B}).$$

Formally $\mathcal{B}$ stands for the subset $\{(X_j, Y_j, Z_j), j = 1, \ldots, n : Z_j = 1\}$. In the following we assume that (X_p, Y_p, Z_p) and (X_q, Y_q, Z_q) are two different observations which do not belong to the set of complete observations $\mathcal{B}$. Consider

$$B_3^2 = E^2(B_2|\mathcal{B})$$

$$
\begin{aligned}
&= E^2\big[(1-Z_p)\{\phi(X_p)-\widetilde{\phi}(X_p)\}|\mathcal{B}\big] \\
&= E\big[(1-Z_p)\{\phi(X_p)-\widetilde{\phi}(X_p)\}|\mathcal{B}\big]\, E\big[(1-Z_q)\{\phi(X_q)-\widetilde{\phi}(X_q)\}|\mathcal{B}\big] \\
&= E\big[(1-Z_p)\{\phi(X_p)-\widetilde{\phi}(X_p)\}(1-Z_q)\{\phi(X_q)-\widetilde{\phi}(X_q)\}|\mathcal{B}\big],
\end{aligned}
$$

where the last equality holds because $\phi(X_p)$ and $\phi(X_q)$ are conditionally independent given $\mathcal{B}$. Then

$$
E(B_3^2) = E\big[(1-Z_1)(1-Z_2)\{\phi(X_1)-\widetilde{\phi}(X_1)\}\,\{\phi(X_2)-\widetilde{\phi}(X_2)\}\big].
$$

This combined with

$$
\begin{aligned}
E(B_2^2) &= \frac{1}{n^2}E\Big(\Big[\sum_{i=1}^{n}(1-Z_i)\{\phi(X_i)-\widetilde{\phi}(X_i)\}\Big]^2\Big) \\
&= \frac{1}{n}E\big[(1-Z)\{\phi(X)-\widetilde{\phi}(X)\}^2\big] \\
&\quad + \frac{n-1}{n}E\big[(1-Z_1)(1-Z_2)\{\phi(X_1)-\widetilde{\phi}(X_1)\}\,\{\phi(X_2)-\widetilde{\phi}(X_2)\}\big]
\end{aligned}
$$

yields

$$
E(B_2^2) - E(B_3^2) = n^{-1}E\big[(1-Z)\{\phi(X)-\widetilde{\phi}(X)\}^2\big] - n^{-1}E(B_3^2). \tag{A.6}
$$

Further we obtain

$$
\begin{aligned}
&E[\{\phi(X)-\widetilde{\phi}(X)\}^2] \\
&= E\Big(\Big[\frac{1}{n}\frac{\sum_{j=1}^{n} Z_j K_b(X_j,X)\{h(X_j,Y_j)-\chi(X)\}}{\pi(X)f(X)}\Big]^2\Big) \\
&\le \frac{c}{n^2}E\Big(\Big[\sum_{j=1}^{n} Z_j K_b(X_j,X)\{h(X_j,Y_j)-\chi(X)\}\Big]^2\Big) \\
&= \frac{c}{n^2}E\Big[\sum_{j=1}^{n} Z_j K_b^2(X_j,X)\{h(X_j,Y_j)-\chi(X)\}^2\Big]
\end{aligned}
$$

$$+\frac{c}{n^2}E\Big[\sum_{i\neq j}^{n} Z_iZ_jK_b(X_i,X)K_b(X_j,X)\{h(X_i,Y_i)-\chi(X)\}\{h(X_j,Y_j)-\chi(X)\}\Big]$$
$$= T_1 + T_2$$

with

$$T_1 = \frac{c}{n}E[Z_1K_b^2(X_1,X)\{h(X_1,Y_1)-\chi(X)\}^2]$$
$$T_2 = \frac{c(n-1)}{n}E[Z_1Z_2K_b(X_1,X)K_b(X_2,X)\{h(X_1,Y_1)-\chi(X)\}\{h(X_2,Y_2)-\chi(X)\}].$$

For T_1 we have

$$\begin{aligned} T_1 &\leq \frac{c}{n}E[\{h(X_1,Y_1)-\chi(X)\}^2] \\ &= \frac{c}{n}[E\{h^2(X_1,Y_1)\}+E\{\chi^2(X)\}-2E\{h(X_1,Y_1)\chi(X)\}] \\ &= \frac{c}{n}[E\{h^2(X_1,Y_1)\}-E^2\{h(X_1,Y_1)\}+E\{\chi^2(X)\}-E^2\{\chi(X)\}] \\ &= \frac{c}{n}\{\mathrm{Var}[h(X,Y)]+\mathrm{Var}[\chi(X)]\}\to 0 \quad (n\to\infty). \end{aligned}$$

In the third step we use

$$E\{h(X_1,Y_1)\chi(X)\} = E\{h(X_1,Y_1)\}E\{\chi(X)\} = E^2\{h(X_1,Y_1)\} = E^2\{\chi(X)\},$$

and in the last statement that the variances are finite by assumption.

The second term T_2 computes to

$$\begin{aligned} T_2 &= \frac{c(n-1)}{n}E\Big(E[Z_1Z_2K_b(X_1,X)K_b(X_2,X)\{h(X_1,Y_1)-\chi(X)\} \\ &\quad \times\{h(X_2,Y_2)\chi(X)\}|X]\Big) \\ &= \frac{c(n-1)}{n}E\Big(E^2[Z_1K_b(X_1,X)\{h(X_1,Y_1)-\chi(X)\}|X]\Big) \\ &\leq c\sup_{x\in\mathcal{I}}E^2[K_b(X_1,x)\{h(X_1,Y_1)-\chi(x)\}] \\ &= c\big(\sup_{x\in\mathcal{I}}\big|E[K_b(X_1,x)\{h(X_1,Y_1)-\chi(x)\}]\big|\big)^2\to 0. \end{aligned}$$

The last step follows from the first conclusion of Lemma A.1. Hence we have

$$E[\{\phi(X) - \widetilde{\phi}(X)\}^2] = T_1 + T_2 \to 0.$$

This combined with (A.6) yields

$$\begin{aligned}
nE\{(B_2 - B_3)^2\} &= n\{E(B_2^2) - 2E(B_2B_3) + E(B_3^2)\} \\
&= n\{E(B_2^2) - 2E\{E(B_2B_3|\mathcal{B})\} + E(B_3^2)\} \\
&= n\{E(B_2^2) - 2E\{B_3E(B_2|\mathcal{B})\} + E(B_3^2)\} \\
&= n\{E(B_2^2) - E(B_3^2)\} \\
&= E[(1-Z)\{\phi(X) - \widetilde{\phi}(X)\}^2] - E(B_3^2) \\
&\leq E[\{\phi(X) - \widetilde{\phi}(X)\}^2] \to 0.
\end{aligned}$$

Now use $E(B_2 - B_3) = 0$ and Chebyshev's inequality to obtain $n^{1/2}|B_2 - B_3| = o_p(1)$. This and (A.5) finally give

$$n^{1/2}|B_1 - B_3| = o_p(1). \tag{A.7}$$

It remains to examine B_3 more closely. Assume that (X_p, Y_p, Z_p) does not belong to the set of complete observations $\mathcal{B}$. We have

$$\begin{aligned}
B_3 &= E(B_2|\mathcal{B}) \\
&= E\Big[\frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\frac{n^{-1}\sum_{j=1}^{n}Z_jK_b(X_j,X_i)\{h(X_j,Y_j)-\chi(X_i)\}}{\pi(X_i)f(X_i)}\Big|\mathcal{B}\Big] \\
&= E\Big[\{1-\pi(X_p)\}\frac{n^{-1}\sum_{j=1}^{n}Z_jK_b(X_j,X_p)\{h(X_j,Y_j)-\chi(X_p)\}}{\pi(X_p)f(X_p)}\Big|\mathcal{B}\Big] \\
&= \frac{1}{n}\sum_{j=1}^{n}Z_jE\Big[\{1-\pi(X_p)\}\frac{K_b(X_j,X_p)\{h(X_j,Y_j)-\chi(X_p)\}}{\pi(X_p)f(X_p)}\Big|\mathcal{B}\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{j=1}^{n} Z_j h(X_j, Y_j) E\Big[\frac{\{1-\pi(X_p)\}K_b(X_j, X_p)}{\pi(X_p)f(X_p)}\Big|X_j\Big] \\
&\qquad -\frac{1}{n}\sum_{j=1}^{n} Z_j E\Big[\frac{\{1-\pi(X_p)\}K_b(X_j, X_p)\chi(X_p)}{\pi(X_p)f(X_p)}\Big|X_j\Big] \\
&= \frac{1}{n}\sum_{j=1}^{n} Z_j h(X_j, Y_j)\frac{1-\pi(X_j)}{\pi(X_j)} - \frac{1}{n}\sum_{j=1}^{n} Z_j \chi(X_j)\frac{1-\pi(X_j)}{\pi(X_j)} + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{j=1}^{n} Z_j \{h(X_j, Y_j) - \chi(X_j)\}\frac{1-\pi(X_j)}{\pi(X_j)} + o_P(n^{-1/2}).
\end{aligned}
$$

The last but one step follows from the first conclusion in Lemma A.1. This combined with (A.4) and (A.7) gives the expansion provided in the theorem:

$$
\begin{aligned}
\widehat{H}_{np} &= A + B_1 \\
&= A + B_3 + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\big[\chi(X_i) + \frac{Z_i}{\pi(X_i)}\{h(X_i, Y_i) - \chi(X_i)\}\big] + o_p(n^{-1/2}).
\end{aligned}
$$

This completes the proof. □

A.3 Proof of Theorem 2.2

Consider $\Gamma = \sum_{i=1}^{n} Z_i g(X_i)\widehat{\varepsilon}_i/n$. We begin with an auxiliary result. A first order Taylor expansion, using Assumption 2.1, yields

$$
\begin{aligned}
&\sum_{i=1}^{n}[g(X_i)\{r_\tau(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^{\mathrm{T}}(\tau-\vartheta)\}]^2 \\
&= \sum_{i=1}^{n} g^2(X_i)\Big[\int_0^1 \{\dot{r}_{\vartheta+u(\tau-\vartheta)}(X_i) - \dot{r}_\vartheta(X_i)\}^{\mathrm{T}}(\tau-\vartheta)\,du\Big]^2 \\
&\le \|\tau-\vartheta\|^2 \sum_{i=1}^{n} g^2(X_i)\int_0^1 \|\dot{r}_{\vartheta+u(\tau-\vartheta)}(X_i) - \dot{r}_\vartheta(X_i)\|^2\,du \\
&\le \|\tau-\vartheta\|^4 \sum_{i=1}^{n} g^2(X_i)L^2.
\end{aligned}
$$

This combined with the square integrability of $\rho_h(X)$, Assumption 2.7, guarantees for any constant c that

$$\sup_{\|\tau-\vartheta\|\le cn^{-1/2}} \sum_{i=1}^{n} \{g(X_i)[r_\tau(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)(\tau-\vartheta)]\}^2 = o_p(1). \tag{A.8}$$

We now approximate Γ by $\sum_{i=1}^n Z_i g(X_i)\varepsilon_i^*/n$, where $\varepsilon_i^* = \varepsilon_i - \dot{r}_\vartheta(X_i)^{\mathrm{T}}(\widehat{\vartheta} - \vartheta)$. We have

$$\begin{aligned} \Big|\frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)\Big| &\le \frac{1}{n}\sum_{i=1}^{n} Z_i |g(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)| \\ &\le \frac{1}{n}\Big\{ n\sum_{i=1}^{n} Z_i g^2(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)^2 \Big\}^{1/2} \\ &= n^{-1/2}\Big\{ \sum_{i=1}^{n} Z_i g^2(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)^2 \Big\}^{1/2}. \end{aligned} \tag{A.9}$$

The second relation uses the Cauchy-Schwarz inequality. Now apply (A.8) to obtain

$$\begin{aligned} \sum_{i=1}^{n} Z_i g^2(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)^2 &= \sum_{i=1}^{n} Z_i g^2(X_i)[\widehat{\varepsilon}_i - \{\varepsilon_i - \dot{r}_\vartheta(X_i)^{\mathrm{T}}(\widehat{\vartheta} - \vartheta)\}]^2 \\ &\le \sum_{i=1}^{n} g^2(X_i)\{r_{\widehat{\vartheta}}(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^{\mathrm{T}}(\widehat{\vartheta} - \vartheta)\}^2 \\ &= o_p(1). \end{aligned} \tag{A.10}$$

This combined with (A.9) gives

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\widehat{\varepsilon}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i^* + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i - \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\dot{r}_\vartheta^{\mathrm{T}}(X_i)(\widehat{\vartheta} - \vartheta) + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i - E\{Zg(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\}(\widehat{\vartheta} - \vartheta) + o_p(n^{-1/2}) \end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i - \frac{1}{n}E\{Zg(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\}I^{-1}\sum_{i=1}^{n} Z_i\sigma^{-2}(X_i)\dot{r}_\vartheta(X_i)\varepsilon_i + o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i\varepsilon_i}{\sigma^2(X_i)}\Big\{\frac{\rho_h(X_i)}{\pi(X_i)} - \dot{r}_\vartheta(X_i)^{\mathrm{T}}I^{-1}\Delta\Big\} + o_p(n^{-1/2}). \tag{A.11}$$

Here we also used the law of large numbers and the fourth equation uses the asymptotic linearity of $\widehat{\vartheta}$ stated in Assumption 2.2. This gives the influence function of the correction term, which involves the unknown quantity $g(x) = \rho_h(x)/\{\pi(x)\sigma^2(x)\}$.

It is left to prove replacing $g(x)$ by a uniformly consistent estimator $\widehat{g}(x)$ does not change the asymptotic expansion, i.e.,

$$\Big|n^{-1}\sum_{i=1}^{n} Z_i\big\{g(X_i) - \widehat{g}(X_i)\big\}\widehat{\varepsilon}_i\Big| = o_p(n^{-1/2}). \tag{A.12}$$

We will use similar arguments as in the prior part of the proof of the theorem, in particular we write again $\varepsilon_i^* = \varepsilon_i - \dot{r}_\vartheta(X_i)^{\mathrm{T}}(\widehat{\vartheta} - \vartheta)$. Using this notation, equation (A.12) becomes

$$\Big|n^{-1}\sum_{i=1}^{n} Z_i\big\{g(X_i) - \widehat{g}(X_i)\big\}\widehat{\varepsilon}_i\Big|$$
$$= \Big|n^{-1}\sum_{i=1}^{n} Z_i\big\{g(X_i) - \widehat{g}(X_i)\big\}\{(\widehat{\varepsilon}_i - \varepsilon_i^*) - \dot{r}_\vartheta(X_i)^{\mathrm{T}}(\widehat{\vartheta} - \vartheta) + \varepsilon_i\}\Big| = o_p(n^{-1/2}).$$

We treat the three parts separately. Analogously as in the proof of (A.9), we obtain for the first part

$$\Big|n^{-1}\sum_{i=1}^{n} Z_i\big\{g(X_i) - \widehat{g}(X_i)\big\}(\widehat{\varepsilon}_i - \varepsilon_i^*)\Big|$$
$$\le n^{-1}\sum_{i=1}^{n}\big|\big\{g(X_i) - \widehat{g}(X_i)\big\}(\widehat{\varepsilon}_i - \varepsilon_i^*)\big|$$
$$\le n^{-1}\Big\{n\sum_{i=1}^{n}\big\{g(X_i) - \widehat{g}(X_i)\big\}^2(\widehat{\varepsilon}_i - \varepsilon_i^*)^2\Big\}^{1/2}$$
$$= n^{-1/2}\sup_{x\in\mathcal{I}}\big\{g(x) - \widehat{g}(x)\big\}^2\Big\{\sum_{i=1}^{n}(\widehat{\varepsilon}_i - \varepsilon_i^*)^2\Big\}^{1/2}$$
$$= o_p(n^{-1/2}) \tag{A.13}$$

In the last step we use the arguments following (A.10) with $g(\cdot) \equiv 1$, and the fact that $\widehat{g}(x)$ is a consistent estimator of $g(x)$. The second part computes to

$$\begin{aligned}
&\Big| n^{-1} \sum_{i=1}^{n} Z_i \big\{ g(X_i) - \widehat{g}(X_i) \big\} \dot{r}_\vartheta(X_i)^{\mathrm{T}} (\widehat{\vartheta} - \vartheta) \Big| \\
&\leq n^{-1} \|\widehat{\vartheta} - \vartheta\| \sum_{i=1}^{n} \big| g(X_i) - \widehat{g}(X_i) \big| \|\dot{r}_\vartheta(X_i)\| \\
&= o_p(n^{-1/2}) \qquad \text{(A.14)}
\end{aligned}$$

where the last step uses Assumptions (R) and (T) as well as the uniform consistency of $\widehat{g}(x)$.

Finally we show

$$n^{-1} \sum_{i=1}^{n} Z_i \big\{ \widehat{g}(X_i) - g(X_i) \big\} \varepsilon_i = o_p(n^{-1/2}). \qquad \text{(A.15)}$$

In equation (A.11), we have seen that $n^{-1} \sum_{i=1}^{n} Z_i g(X_i) \varepsilon_i$ is part of the approximation and therefore of order $O_p(n^{-1/2})$. The term on the left-hand side of (A.15) is approximately conditionally centered (given X_i). Since $\widehat{g}(x) - g(x)$ is asympotically negligible, we obtain the desired order $o_p(n^{-1/2})$.

Combining (A.13), (A.14) and (A.15) gives the desired statement

$$\Big| n^{-1} \sum_{i=1}^{n} Z_i \big\{ g(X_i) - \widehat{g}(X_i) \big\} \widehat{\varepsilon}_i \Big| = o_p(n^{-1/2}).$$

To prove that the term in (A.15) is exactly conditionally centered, we propose using leave-one out estimators $\widetilde{\rho}(X_i)$, $\widetilde{\sigma}^2(X_i)$ and $\widetilde{\pi}(X_i)$ to estimate $\widehat{g}(X_i)$ ($i = 1, \ldots, n$), i.e.,

$$\widehat{g}(X_i) = \frac{\widetilde{\rho}(X_i)}{\widetilde{\sigma}^2(X_i) \widetilde{\pi}(X_i)}.$$

Choose, for example,

$$\widetilde{\sigma}^2(X_i) = \frac{\sum_{j=1,j\neq i}^n Z_j K_b(X_i - X_j)\{Y_j - r_{\widetilde{\vartheta}_i}(X_j)\}^2}{\sum_{j=1,j\neq i}^n Z_j K_b(X_i - X_j)}$$

where $\widetilde{\vartheta}_i$ is some consistent estimator of ϑ that does not use (X_i, Y_i) if that pair is observed. The other two leave-one-out estimator are defined similarly. Thanks to this construction $\widehat{g}(X_i)$ is independent of Y_i and Z_i conditional on X_i, and we obtain (suppressing the subscript i)

$$\begin{aligned} &E\big\{Z(\widehat{g}(X) - g(X))\varepsilon\big\} = E\big\{Z\widehat{g}(X)\varepsilon\big\} \\ &= E\Big\{E\{Z\widehat{g}(X)\varepsilon|X\}\Big\} = E\Big\{\pi(X)E\{\widehat{g}(X)|X\}E\{\varepsilon|X\}\Big\} = 0. \end{aligned}$$

□

A.4 Proof of Theorem 2.3

Müller et al. (2006) and Müller (2009) show that the canonical gradient $g_*(X, ZY, Z)$ from (2.4), now specified for the functional $E\{h(X,Y)\}$, is determined by

$$\begin{aligned} &E\{u_*(X)u(X)\} + E\{Zv_*(X,Y)v(X,Y)\} \\ &+E[\{Z - \pi(X)\}^2 w_*(X)w(X)] = E[h(X,Y)\{u(x) + v(X,Y)\}] \end{aligned} \tag{A.16}$$

for all $u \in U$, $v \in V$ and $w \in W$. Here we use the fact that the canonical gradient of $E\{h(X,Y)\}$ is a projection of a gradient of $E\{h(X,Y)\}$ onto the tangent space. To determine the specific form of g_* we set $u = 0$ and $v = 0$ in (A.16), which gives

$$w_* = 0. \tag{A.17}$$

Then, setting $v = 0$ in (A.16) yields that $u_*(X)$ is the projection of $h(X,Y)$ onto U:

$$u_*(X) = E\{h(X,Y)|X\} - E\{h(X,Y)\} = \chi(X) - E\{h(X,Y)\}. \tag{A.18}$$

In order to find v_* we must take the parametric model structure into account, i.e., the special form of the subset $V \subset V_0$. To specify the tangent space V concerning the conditional distribution, we introduce perturbations s and t of the two parameters $f(\cdot|x)$ and ϑ. Write $F(\cdot|x)$ for the conditional distribution function of $f(\cdot|x)$ and assume that $f(\cdot|x)$ has finite Fisher information for location, $E\ell^2(\varepsilon|x) < \infty$, where $\ell(\cdot|x) = -f'(\cdot|x)/f(\cdot|x)$ is the score function. The perturbed conditional distribution is

$$Q_{nv}(x, dy) = Q_{nsa}(x, dy) = f_{ns}\{y - r_{\vartheta_{na}(x)}|x\}dy$$

with $\vartheta_{na} = \vartheta + n^{-1/2}a$, $a \in \mathbb{R}^d$, $f_{ns}(y|x) = f(y|x)\{1 + n^{-1/2}s(x, y)\}$ and $s \in S$, where

$$S = \big\{s \in L_2(F) : \int s(x, y)f(y|x)dy = 0, \int ys(x, y)f(y|x)dy = 0\big\}.$$

Here S is determined by two constraints: the perturbed error conditional density $f_{ns}(\cdot|x)$ must integrate to 1, $\int f_{ns}(y|x)dy = 1$, and must be centered at zero, $\int yf_{ns}(y|x)dy = 0$. As in Schick (1993), Section 3, we have

$$\begin{aligned}
&f_{ns}\{y - r_{\vartheta_{na}}(x)|x\} \\
&= f\{y - r_{\vartheta_{na}}(x)|x\}[1 + n^{-1/2}s\{x, y - r_{\vartheta_{na}}(x)\}] \\
&\doteq [f\{y - r_\vartheta(x)|x\} - n^{-1/2}f'\{y - r_\vartheta(x)|x\}\dot{r}_\vartheta(x)^{\mathrm{T}}a][1 + n^{-1/2}s\{x, y - r_\vartheta(x)\}] \\
&\doteq f\{y - r_\vartheta(x)|x\}\Big(1 + n^{-1/2}\Big[s\{x, y - r_\vartheta(x)\} - \frac{f'\{y - r_\vartheta(x)|x\}}{f\{y - r_\vartheta(x)|x\}}\dot{r}_\vartheta(x)^{\mathrm{T}}a\Big]\Big) \\
&= f\{y - r_\vartheta(x)|x\}(1 + n^{-1/2}[s\{x, y - r_\vartheta(x)\} + \ell\{y - r_\vartheta(x)|x\}\dot{r}_\vartheta(x)^{\mathrm{T}}a]).
\end{aligned}$$

Therefore the subspace V of V_0 is

$$V = \big\{s\{x, y - r_\vartheta(x)\} + \ell\{y - r_\vartheta(x)|x\}\dot{r}_\vartheta(x)^{\mathrm{T}}a : s \in S, a \in \mathbb{R}^d\big\}.$$

Setting $\widetilde{V} = \{v(X,Y) : v \in V\}$ and writing $v \in \widetilde{V}$ as a sum of three terms, we obtain

$$\begin{aligned} v(X,Y) &= s(X,\varepsilon) + \ell(\varepsilon|X)\dot{r}_\vartheta(X)^{\mathrm{T}}a \\ &= s(X,\varepsilon) + \Big\{\ell(\varepsilon|X) - \frac{\varepsilon}{\sigma^2(X)}\Big\}\dot{r}_\vartheta(X)^{\mathrm{T}}a + \frac{\varepsilon}{\sigma^2(X)}\dot{r}_\vartheta(X)^{\mathrm{T}}a. \end{aligned}$$

The third term is obviously an element of

$$V_1 = \{\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon : a \in \mathbb{R}^d\}.$$

It is easy to check that the first two terms (and their sum) belong to

$$V_2 = \{t(X,Y) : t \in S\}$$

and that V_1 and V_2 are orthogonal. Hence we can write $\widetilde{V}$ as an orthogonal sum, $\widetilde{V} = V_1 \oplus V_2$. To specify v_* in the canonical gradient formula (2.4), we use this presentation and write

$$\begin{aligned} v_*(X,Y) &= \sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a_*\varepsilon + t_*(X,Y), \\ v(X,Y) &= \sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon + t(X,Y), \end{aligned} \tag{A.19}$$

where $a_*,\ a \in \mathbb{R}^d$ and $t_*, t \in S$. Setting $u = 0$ and $w = 0$ in equation (A.16), we obtain

$$\begin{aligned} &E[Z\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a_*\varepsilon + t_*\}\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon + t\}] \\ &= E[h(X,Y)\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon + t\}]. \end{aligned} \tag{A.20}$$

Set $t = 0$ in (A.20) and use $E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon t_*\} = 0$, which holds since $t_* \in S$. Then (A.20) becomes

$$E\{Z\sigma^{-4}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a_*\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon^2\} = E\{h(X,Y)\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}}a\varepsilon\},$$

and, since the equation must be satisfied for arbitrary vectors a,

$$a_*^{\mathrm{T}} E\{Z\sigma^{-4}(X)\varepsilon^2 \dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\} = E\{h(X,Y)\sigma^{-2}(X)\varepsilon \dot{r}_\vartheta(X)^{\mathrm{T}}\}.$$

The term on the left-hand side computes to

$$\begin{aligned} a_*^{\mathrm{T}} E\{Z\sigma^{-4}(X)\varepsilon^2 \dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\} &= a_*^{\mathrm{T}} E[E\{Z\sigma^{-4}(X)\varepsilon^2 \dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}|X\}] \\ &= a_*^{\mathrm{T}} E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\}, \end{aligned}$$

and, assuming $E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\}$ is invertible, we obtain

$$a_* = [E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^{\mathrm{T}}\}]^{-1} E\{h(X,Y)\sigma^{-2}(X)\varepsilon \dot{r}_\vartheta(X)\} = I^{-1}\Delta, \tag{A.21}$$

with I and Δ as in Theorem 2.2 and Corollary 2.1. Now set $a = 0$ in (A.20) and use

$$\begin{aligned} E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}} a_* \varepsilon t\} &= E[E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}} a_* \varepsilon t|X\}] \\ &= E\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^{\mathrm{T}} a_* \pi(X) E(\varepsilon t|X)\} = 0 \end{aligned}$$

to obtain

$$E(Zt_*t) = E\{h(X,Y)t\}.$$

Writing this as an iterated expectation,

$$E\{E(Zt_*t|X)\} = E\{\pi(X)E(t_*t|X)\} = E[E\{h(X,Y)t|X\}], \tag{A.22}$$

we see that $h(X,Y)/\pi(X)$ is a candidate for $t_*(X,Y)$. Since t_* must be in S, we choose a suitably

modified version, namely

$$\begin{aligned} t_*(X,Y) &= \frac{1}{\pi(X)}[h(X,Y) - E\{h(X,Y)|X\} - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)\varepsilon] \\ &= \frac{1}{\pi(X)}\Big\{h(X,Y) - \chi(X) - \frac{\varepsilon\rho_h(X)}{\sigma^2(X)}\Big\} \end{aligned} \tag{A.23}$$

with $\rho_h(x) = E\{h(X,Y)\varepsilon|X = x\}$. To verify (A.23) formally, we show that t_* satisfies characterization (A.22) and that t_* is in $V_2 = \{t(X,Y) : t \in S\}$. To prove the first part we consider $t = t(X,Y) \in V_2$, that is, by definition of S, $E(t|X) = 0$ and $E(t\varepsilon|X) = 0$. Then

$$\begin{aligned} &E(Zt_*t|X) \\ &= E\big(\big[h(X,Y) - E\{h(X,Y)|X\} - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)\varepsilon\big]t|X\big) \\ &= E\{h(X,Y)t\} - E\{h(X,Y)|X\}E(t|X) - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)E(t\varepsilon|X) \\ &= E\{h(X,Y)t\}, \end{aligned}$$

which shows that t_* satisfies (A.22). The second part, $t_* = t_*(X,Y) \in V_2$, follows from

$$\begin{aligned} E(t_*|X) = \frac{1}{\pi(X)}[&E\{h(X,Y)|X\} - E\{h(X,Y)|X\} \\ &- E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)E(\varepsilon|X)] = 0 \end{aligned}$$

and

$$\begin{aligned} E(t_*\varepsilon|X) &= \frac{1}{\pi(X)}\big[E\{h(X,Y)\varepsilon|X\} - E\{h(X,Y)|X\}E(\varepsilon|X) \\ &\qquad - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)E(\varepsilon^2|X)\big] \\ &= \frac{1}{\pi(X)}\big[E\{h(X,Y)\varepsilon|X\} - E\{h(X,Y)\varepsilon|X\}\big] = 0. \end{aligned}$$

Therefore (A.23) is true. Now plug (A.21) and (A.23) into the formula for v_* in (A.19) to obtain

$$v_*(X,Y) = \frac{\varepsilon}{\sigma^2(X)}\dot{r}_\vartheta(X)^{\mathrm{T}} I^{-1}\Delta + \frac{1}{\pi(X)}\Big\{h(X,Y) - \chi(X) - \frac{\varepsilon\rho_h(X)}{\sigma^2(X)}\Big\}. \tag{A.24}$$

Combining (2.4), (A.17), (A.18) and (A.24) yields the canonical gradient of $E\{h(X,Y)\}$ given in the theorem. □

APPENDIX B

TECHNICAL DETAILS OF CHAPTER 3

B.1 Auxiliary lemmas

Lemma B.1. *Use the notations from Section 3.1 and write*

$$\widetilde{\beta}_{nkm} = n^{1/2}(X_{k\cdot a}^{\mathrm{T}} B_{nkm} X_{k\cdot a})^{-1} X_{k\cdot a}^{\mathrm{T}} \psi_{nkm}(\varepsilon)$$

for $k = 1, \ldots, K$ *and* $m = 1, \ldots, M$. *Then provided Assumptions 3.1, 3.2, 3.3 and 3.4 are satisfied, we have*

$$\|\widetilde{\beta}_{nkm}\| = O_p((q_n \log n)^{1/2}).$$

Proof of Lemma B.1: We calculate

$$\begin{aligned}
\|\widetilde{\beta}_{nkm}\|^2 &= n\psi_{nkm}(\varepsilon)^{\mathrm{T}} X_{k\cdot a}(X_{k\cdot a}^{\mathrm{T}} B_{nkm} X_{k\cdot a})^{-2} X_{k\cdot a}^{\mathrm{T}} \psi_{nkm}(\varepsilon) \\
&\leq \lambda_{\min}(n^{-1} X_{k\cdot a}^{\mathrm{T}} B_{nkm} X_{k\cdot a})^{-2} n^{-1} \psi_{nkm}(\varepsilon)^{\mathrm{T}} X_{k\cdot a} X_{k\cdot a}^{\mathrm{T}} \psi_{nkm}(\varepsilon) \\
&\leq cn^{-1} \psi_{nkm}(\varepsilon)^{\mathrm{T}} X_{k\cdot a} X_{k\cdot a}^{\mathrm{T}} \psi_{nkm}(\varepsilon) \\
&\leq cn^{-1} q_n \Big(\max_{1\leq j\leq q_n} |\psi_{nkm}(\varepsilon)^{\mathrm{T}} X_{k\cdot j}| \Big)^2 \\
&= cn^{-1} q_n \Big(\max_{1\leq j\leq q_n} \Big| \sum_{i=1}^{n} \psi_{kmi}(\varepsilon) X_{kij} \Big| \Big)^2 \qquad \text{(B.1)}
\end{aligned}$$

where the third step uses Assumption 3.2 and 3.3. Since $\psi_{kmi}(\varepsilon)X_{kij}$ has mean zero and is bounded by Assumption 3.1, Hoeffding's inequality gives

$$\mathrm{pr}\Big\{ \Big| \sum_{i=1}^{n} \psi_{kmi}(\varepsilon) X_{kij} \Big| \geq L_n (n \log n)^{1/2} \Big\} \leq 2 \exp\{-cL_n^2 \log n\}.$$

for any positive sequence $L_n \to \infty$. It follows that

$$\begin{aligned}
&\mathrm{pr}\Big\{ \max_{1\le j\le q_n} \Big| \sum_{i=1}^{n} \psi_{kmi}(\varepsilon) X_{kij} \Big| \ge L_n (n \log n)^{1/2} \Big\} \\
&\le \sum_{j=1}^{q_n} \mathrm{pr}\Big\{ \Big| \sum_{i=1}^{n} \psi_{kmi}(\varepsilon) X_{kij} \Big| \ge L_n (n \log n)^{1/2} \Big\} \\
&\le 2 q_n \exp\{-C L_n^2 \log n\} = 2 q_n n^{-C L_n^2} \to 0. \qquad \text{(B.2)}
\end{aligned}$$

where the last step holds true because $q_n = o(n^{1/2})$; see Assumption 3.4. Therefore

$$\max_{1\le j\le q_n} \Big| \sum_{i=1}^{n} \psi_{kmi}(\varepsilon) X_{kij} \Big| = O_p((n \log n)^{1/2}).$$

This combined with (B.1) gives $\|\widetilde{\beta}_{nkm}\|^2 = O_p(q_n \log n)$, which completes the proof. □

Lemma B.2. *Set $\mathcal{M}_1^* = \{\mathcal{D} : \mathcal{D} \in \mathcal{M}, \mathcal{D}^* \subset \mathcal{D}\}$ and use the notations from Section 3.2. Let Assumptions 3.1, 3.3, 3.6 and 3.7 be satisfied. Let c_4 be the constant from Assumption 3.7. Then we have, for $k = 1, \ldots, K$, $m = 1, \ldots, M$, and any positive sequence L_n satisfying $L_n \to \infty$ and $1 \le L_n (\log n)^{1/2} \le n^{1/10 - c_4/5}$,*

$$\mathrm{pr}\Big\{ \Big| \sum_{i=1}^{n} \{\rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{km\mathcal{D}}) - \rho_m(\varepsilon_{kmi})\} \Big| \le L_n |\mathcal{D}| \log n, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} \to 1.$$

Proof of Lemma B.2: Under Assumptions 3.1, 3.3, 3.6 and 3.7, Lemma A.2 in the supplement to Lee et al. (2014) gives

$$\lim_{L\to\infty} \lim_{n\to\infty} \mathrm{pr}\Big\{ \|\widehat{\theta}_{km\mathcal{D}} - \theta^*_{km\mathcal{D}}\| \le L n^{-1/2} (|\mathcal{D}| \log p_n)^{1/2}, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} = 1. \qquad \text{(B.3)}$$

Then, as $L_n \to \infty$,

$$\mathrm{pr}\Big\{ \|\widehat{\theta}_{km\mathcal{D}} - \theta^*_{km\mathcal{D}}\| \le L_n n^{-1/2} (|\mathcal{D}| \log p_n)^{1/2}, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} \to 1. \qquad \text{(B.4)}$$

Under Assumptions 3.1, 3.3, 3.6 and 3.7, and since $1 \le L_n (\log n)^{1/2} \le n^{1/10 - c_4/5}$, we can apply

Lemma A.1 in the supplement to Lee et al. (2014), which implies

$$\max_{\mathcal{D}\in\mathcal{M}_1^*}\Big||\mathcal{D}|^{-1}\Big[\widehat{V}_{km\mathcal{D}}-E(\widehat{V}_{km\mathcal{D}}|X_{k\cdot\mathcal{D}})+2\sum_{i=1}^{n}X_{ki\mathcal{D}}^{\mathrm{T}}(\widehat{\theta}_{km\mathcal{D}}-\theta_{km\mathcal{D}}^*)\psi_{kmi}(\varepsilon)\Big]\Big|=o_p(1) \tag{B.5}$$

with $\widehat{V}_{km\mathcal{D}}=\sum_{i=1}^{n}\{\rho_m(Y_{ki}-X_{ki\mathcal{D}}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}})-\rho_m(\varepsilon_{kmi})\}$. Then we have, on an event with probability tending to one,

$$\begin{aligned}
&\Big|\sum_{i=1}^{n}X_{ki\mathcal{D}}^{\mathrm{T}}(\widehat{\theta}_{km\mathcal{D}}-\theta_{km\mathcal{D}}^*)\psi_{kmi}(\varepsilon)\Big|\\
&\le\|\widehat{\theta}_{km\mathcal{D}}-\theta_{km\mathcal{D}}^*\|\Big\|\sum_{i=1}^{n}X_{ki\mathcal{D}}\psi_{kmi}(\varepsilon)\Big\|\\
&\le\|\widehat{\theta}_{km\mathcal{D}}-\theta_{km\mathcal{D}}^*\||\mathcal{D}|^{1/2}\max_{1\le j\le p_n}\Big|\sum_{i=1}^{n}X_{kij}\psi_{kmi}(\varepsilon)\Big|\\
&\le L_n n^{-1/2}(|\mathcal{D}|\log p_n)^{1/2}|\mathcal{D}|^{1/2}L_n(n\log n)^{1/2}=L_n^2|\mathcal{D}|\log n
\end{aligned} \tag{B.6}$$

for any $\mathcal{D}\in\mathcal{M}_1^*$. The last but one step uses (B.2) and (B.4). From Assumption 3.7 we have $p_n=O(n^{c_3})$. Hence (B.2) holds true when q_n is substituted by p_n. We also have, for any $\theta_{\mathcal{D}}\in\mathbb{R}^{|\mathcal{D}|}$ satisfying $\|\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*\|\le L_n n^{-1/2}(|\mathcal{D}|\log p_n)^{1/2}$,

$$\begin{aligned}
&\Big|\sum_{i=1}^{n}E\{\rho_m(Y_{ki}-X_{ki\mathcal{D}}^{\mathrm{T}}\theta_{\mathcal{D}})-\rho_m(\varepsilon_{kmi})|X_{ki}\}\Big|\\
&=\sum_{i=1}^{n}E\Big\{\int_0^{X_{ki\mathcal{D}}^{\mathrm{T}}(\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*)}I(\varepsilon_{kmi}\le s)-I(\varepsilon_{kmi}\le 0)\,ds\Big|X_{ki}\Big\}\\
&=\sum_{i=1}^{n}\int_0^{X_{ki\mathcal{D}}^{\mathrm{T}}(\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*)}F_{km}(s|X_{ki})-F_{km}(0|X_{ki})\,ds\\
&=\sum_{i=1}^{n}\int_0^{X_{ki\mathcal{D}}^{\mathrm{T}}(\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*)}sf_{km}(\bar{s}|X_{ki})\,ds\\
&\le c(\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*)^{\mathrm{T}}\sum_{i=1}^{n}(X_{ki\mathcal{D}}X_{ki\mathcal{D}}^{\mathrm{T}})(\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*)\\
&\le cn\lambda_{\max}(n^{-1}X_{k\cdot\mathcal{D}}^{\mathrm{T}}X_{k\cdot\mathcal{D}})\|\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*\|^2\\
&\le cn\|\theta_{\mathcal{D}}-\theta_{km\mathcal{D}}^*\|^2\le CL_n^2|\mathcal{D}|\log p_n.
\end{aligned} \tag{B.7}$$

The first step in the above results from Knight's identity (Knight (1998)). In the second step, $F_{km}(\cdot|X_k)$ is the conditional distribution function of ε_{km} given X_k. The third step uses a Taylor expansion with some $\bar{s}$ between 0 and $X_{ki\mathcal{D}}^{\mathrm{T}}(\theta_{\mathcal{D}} - \theta_{km\mathcal{D}}^*)$. The fourth step holds true because of Assumption 3.3 and the fact that

$$\begin{aligned}\sup_{1\leq i\leq n}|X_{ki\mathcal{D}}^{\mathrm{T}}(\theta_{\mathcal{D}} - \theta_{km\mathcal{D}}^*)| &\leq \sup_{1\leq i\leq n}\|X_{ki\mathcal{D}}\|\|\theta_{\mathcal{D}} - \theta_{km\mathcal{D}}^*\| \\ &\leq cL_n d_n n^{-1/2}(\log n)^{1/2} \leq cn^{4c_4/5-2/5}(\log n)^{1/2} \to 0\end{aligned}$$

(Assumptions 3.1 and 3.7). Combining (B.4), (B.5), (B.6) and (B.7) yields

$$\begin{aligned}\widehat{V}_{km\mathcal{D}} &\leq |E(\widehat{V}_{km\mathcal{D}}|X_{k\cdot\mathcal{D}})| + 2\Big|\sum_{i=1}^{n} X_{ki\mathcal{D}}^{\mathrm{T}}(\widehat{\theta}_{km\mathcal{D}} - \theta_{km\mathcal{D}}^*)\psi_{kmi}(\varepsilon)\Big| + |\mathcal{D}|o_p(1) \\ &\leq cL_n^2|\mathcal{D}|\log p_n + L_n^2|\mathcal{D}|\log n + |\mathcal{D}|o_p(1) \\ &\leq cL_n^2|\mathcal{D}|\log n\end{aligned}$$

for any $\mathcal{D} \in \mathcal{M}_1^*$, with probability approaching one, where the $o_p(1)$ is the term from (B.5). This finishes the proof. □

B.2 Proof of Theorem 3.1

Under Assumptions 3.1-3.4, Lemma 6 of Sherwood & Wang (2016) gives

$$\|n^{1/2}(\widehat{\theta}_{km} - \theta_{km}^*) - \widetilde{\beta}_{nkm}\| = o_p(1) \tag{B.8}$$

for every k and m, with $\widetilde{\beta}_{nkm}$ defined in Lemma B.1. Therefore

$$\|\widehat{\theta}_{km} - \theta_{km}^*\| = O_p(n^{-1/2}(q_n \log n)^{1/2}) \tag{B.9}$$

It follows that for every k and m,

$$\max_{1\leq j\leq q_n}|\widehat{\theta}_{kmj} - \theta_{kmj}^*| \leq \|\widehat{\theta}_k - \theta_k^*\| = O_p(n^{-1/2}(q_n \log n)^{1/2}) = O_p(n^{(c_1-1)/2}(\log n)^{1/2}),$$

therefore

$$\max_{1\le j\le q_n}\|\widehat{\theta}^{(j)}-\theta^{*(j)}\|_1 \le KM\max_{1\le k\le K}\max_{1\le m\le M}\max_{1\le j\le q_n}|\widehat{\theta}_{kmj}-\theta^*_{kmj}| = O_p(n^{(c_1-1)/2}(\log n)^{1/2}).$$

which, combined with Assumption 3.5, yields

$$\begin{aligned}\min_{1\le j\le q_n}\|\widehat{\theta}^{(j)}\|_1 &\ge \min_{1\le j\le q_n}\|\theta^{*(j)}\|_1 - \max_{1\le j\le q_n}\|\widehat{\theta}^{(j)}-\theta^{*(j)}\|_1\\ &\ge Cn^{(c_2-1)/2} - \{n^{(c_1-1)/2}(\log n)^{1/2}\} = O_p(n^{(c_2-1)/2})\end{aligned}$$

We assume $\lambda_n = o(n^{(c_2-1)/2})$, which implies

$$\mathrm{pr}\Big\{\min_{1\le j\le q_n}\|\widehat{\theta}^{(j)}\|_1 \ge a\lambda_n\Big\}\to 1. \tag{B.10}$$

Consider the subderivative of the objective function (3.2) with respect to $\theta^{(j)}$ is

$$\frac{\partial\Gamma_{\lambda_n}(\theta)}{\partial\theta^{(j)}} = \begin{cases}\partial\ell_n(\theta)/\partial\theta^{(j)} + \lambda_n\mathbb{S}(\theta^{(j)}), & \|\theta^{(j)}\|_1\le\lambda_n,\\ \partial\ell_n(\theta)/\partial\theta^{(j)} + \mathbb{S}(\theta^{(j)})(a\lambda_n-\|\theta^{(j)}\|_1)/(a-1), & \lambda_n<\|\theta^{(j)}\|_1<a\lambda_n,\\ \partial\ell_n(\theta)/\partial\theta^{(j)}, & a\lambda_n\le\|\theta^{(j)}\|_1,\end{cases} \tag{B.11}$$

where $\mathbb{S}(\theta^{(j)}) = (\mathrm{Sign}(\theta_{11j}),\ldots,\mathrm{Sign}(\theta_{1Mj}),\ldots,\mathrm{Sign}(\theta_{K1j}),\ldots,\mathrm{Sign}(\theta_{KMj}))^{\mathrm{T}}$ with $\mathrm{Sign}(x) = x/|x|$ for $x\neq 0$, and $\mathrm{Sign}(x) = \{u : |u|\le 1\}$ for $x = 0$. Thus (B.10) implies that, with probability tending to one, $\widehat{\theta}^{(j)}$ $(1\le j\le q_n)$ belongs to the third case in (B.11), and

$$0\in\frac{\partial\ell(\theta)}{\partial\theta^{(j)}}\Big|_{\theta=\widehat{\theta}} = \frac{\partial\Gamma_{\lambda_n}(\theta)}{\partial\theta^{(j)}}\Big|_{\theta=\widehat{\theta}} \tag{B.12}$$

because $\widehat{\theta}$ is a local minimizer of $\ell_n(\theta)$.

Under Assumptions 3.1-3.5, Lemma 2.3 of Wang et al. (2012) gives that for every k and m,

$$\mathrm{pr}\Big\{ \max_{q_n<j\le p_n} \Big| \frac{\partial \ell(\theta)}{\partial \theta_{kmj}}\Big|_{\theta=\widehat{\theta}} \Big| > \lambda_n \Big\} \to 0. \tag{B.13}$$

Since $\|\widehat{\theta}^{(j)}\|_1 = 0$ for $q_n < j \le p_n$, which belongs to the first case in (B.11), we have

$$\frac{\partial \Gamma_{\lambda_n}(\theta)}{\partial \theta^{(j)}}\Big|_{\theta=\widehat{\theta}} = \frac{\partial \ell(\theta)}{\partial \theta^{(j)}}\Big|_{\theta=\widehat{\theta}} + \lambda_n \mathbb{S}(\mathbf{0}) \tag{B.14}$$

Since $\mathbb{S}(\mathbf{0}) = \{(u_1, \ldots, u_K) : |u_k| \le 1, k = 1 \ldots, K\}$, (B.13) and (B.14) imply that for $q_n < j \le p_n$,

$$\mathrm{pr}\Big\{ 0 \in \frac{\partial \Gamma_{\lambda_n}(\theta)}{\partial \theta^{(j)}}\Big|_{\theta=\widehat{\theta}} \Big\} \to 1. \tag{B.15}$$

Combing (B.12) and (B.15) completes the proof. □

B.3 Proof of Theorem 3.2

Set $\widehat{\beta}_n = n^{1/2}(\widehat{\theta}_a - \theta_a^*)$ and $\widetilde{\beta}_n = n^{-1/2} R_n^{-1} X_a^{\mathrm{T}} \psi_n(\varepsilon)$, and write

$$A_n \Sigma_n^{-1/2} \widetilde{\beta}_n = \sum_{i=1}^{n} D_{ni}$$

where $D_{ni} = n^{-1/2} A_n \Sigma_n^{-1/2} R_n^{-1} \delta_{ni}$, $\delta_{ni} = (\psi_{1\cdot i}(\varepsilon)^{\mathrm{T}} \otimes X_{1ia}^{\mathrm{T}}, \ldots, \psi_{K\cdot i}(\varepsilon)^{\mathrm{T}} \otimes X_{Kia}^{\mathrm{T}})^{\mathrm{T}}$ and $\psi_{k\cdot i}(\varepsilon) = (\psi_{k1i}(\varepsilon), \ldots, \psi_{kMi}(\varepsilon))^{\mathrm{T}}$ for every k and i. We have $E(D_{ni}) = \mathbf{0}$ since $E(\delta_{ni}) = \mathbf{0}$, and

$$\begin{aligned}
\sum_{i=1}^{n} E(D_{ni} D_{ni}^{\mathrm{T}}) &= n^{-1} E\Big[A_n \Sigma_n^{-1/2} R_n^{-1} \Big\{ \sum_{i=1}^{n} E(\delta_{ni}\delta_{ni}^{\mathrm{T}} | \mathcal{X}) \Big\} R_n^{-1} \Sigma_n^{-1/2} A_n^{\mathrm{T}} \Big] \\
&= E\{ A_n \Sigma_n^{-1/2} R_n^{-1} (n^{-1} X_a^{\mathrm{T}} H_n X_a) R_n^{-1} \Sigma_n^{-1/2} A_n^{\mathrm{T}} \} \\
&= E(A_n \Sigma_n^{-1/2} R_n^{-1} S_n R_n^{-1} \Sigma_n^{-1/2} A_n^{\mathrm{T}}) = A_n A_n^{\mathrm{T}} \to G
\end{aligned}$$

We have, for any $\eta > 0$,

$$\begin{aligned}
\sum_{i=1}^{n} E\{\|D_{ni}\|^2 I(\|D_{ni}\| > \eta)\} &\leq \eta^{-2} \sum_{i=1}^{n} E(\|D_{ni}\|^4) \\
&= (n\eta)^{-2} \sum_{i=1}^{n} E\{(\delta_{ni}^{\mathrm{T}} R_n^{-1} \Sigma_n^{-1/2} A_n^{\mathrm{T}} A_n \Sigma_n^{-1/2} R_n^{-1} \delta_{ni})^2\} \\
&\leq (n\eta)^{-2} \lambda_{\max}^2 (A_n^{\mathrm{T}} A_n) \sum_{i=1}^{n} E\{(\delta_{ni}^{\mathrm{T}} R_n^{-1} \Sigma_n^{-1} R_n^{-1} \delta_{ni})^2\} \\
&\leq cn^{-2} \sum_{i=1}^{n} E\{(\delta_{ni}^{\mathrm{T}} S_n^{-1} \delta_{ni})^2\} \\
&\leq cn^{-2} \sum_{i=1}^{n} E\{\lambda_{\min}(S_n)^{-2} \|\delta_{ni}\|^4\} \\
&\leq cn^{-2} \sum_{i=1}^{n} E(\|\delta_{ni}\|^4) \\
&= cn^{-2} \sum_{i=1}^{n} E\Big\{ \Big(\sum_{k=1}^{K} \sum_{m=1}^{M} \psi_{kmi}^2 \|X_{kia}\|^2 \Big)^2 \Big\} \\
&\leq cn^{-2} \sum_{i=1}^{n} E\Big\{ \Big(\max_{1 \leq k \leq K} \|X_{kia}\| \Big)^4 \Big\} \\
&\leq cn^{-1} E\Big\{ \Big(\max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \|X_{kia}\| \Big)^4 \Big\} \\
&\leq cn^{-1} q_n^2 = o(1)
\end{aligned}$$

with $\lambda_{\max}(\cdot)$ being the largest eigenvalue of a square matrix. The fourth step in the above results from the fact that $\lambda_{\max}(A_n^{\mathrm{T}} A_n) \to C$. The sixth step uses the condition that $\lambda_{\min}(S_n)$ is uniformly bounded away from zero. The last but one step holds true because of Assumption 3.1, and the last step uses Assumption 3.4. This shows that the Lindeberg-Feller condition for the central limit theorem is satisfied, i.e., we obtain

$$A_n \Sigma_n^{-1/2} \widetilde{\beta}_n = \sum_{i=1}^{n} D_{ni} \to N(0, G) \text{ in distribution } (n \to \infty). \tag{B.16}$$

It is obvious that $\widetilde{\beta}_n = (\widetilde{\beta}_{n11}^{\mathrm{T}}, \ldots, \widetilde{\beta}_{n1M}^{\mathrm{T}}, \ldots, \widetilde{\beta}_{nK1}^{\mathrm{T}}, \ldots, \widetilde{\beta}_{nKM}^{\mathrm{T}})^{\mathrm{T}}$ with $\widetilde{\beta}_{nkm}$ defined in Lemma B.1. Hence, using (B.8), we have

$$\|\widehat{\beta}_n - \widetilde{\beta}_n\| \leq \sum_{k=1}^{K}\sum_{m=1}^{M} \|\widehat{\beta}_{nkm} - \widetilde{\beta}_{nkm}\| = o_p(1).$$

It follows that

$$\begin{aligned} \|A_n\Sigma_n^{-1/2}(\widehat{\beta}_n - \widetilde{\beta}_n)\|^2 &= (\widehat{\beta}_n - \widetilde{\beta}_n)^{\mathrm{T}}\Sigma_n^{-1/2}A_nA_n^{\mathrm{T}}\Sigma_n^{-1/2}(\widehat{\beta}_n - \widetilde{\beta}_n) \\ &\leq \lambda_{\max}(A_nA_n^{\mathrm{T}})\lambda_{\min}(\Sigma_n)^{-1}\|\widehat{\beta}_n - \widetilde{\beta}_n\|^2 = o_p(1). \end{aligned}$$

In the last step we have used $\lambda_{\max}(A_nA_n^{\mathrm{T}}) \to C$, Assumption 3.2 and the condition that $\lambda_{\min}(S_n)$ is uniformly bounded away from zero. This combined with (B.16) yields

$$n^{1/2}A_n\Sigma_n^{-1/2}(\widehat{\theta}_a - \theta_a^*) = A_n\Sigma_n^{-1/2}\widehat{\beta}_n \to N(0, G) \text{ in distribution } (n \to \infty).$$

□

B.4 Proof of Theorem 3.3

Denote the set of overfitted models $\mathcal{M}_1 = \{\mathcal{D} \in \mathcal{M} : \mathcal{D}^* \subset \mathcal{D}, \mathcal{D} \neq \mathcal{D}^*\}$ and the set of underfitted models $\mathcal{M}_2 = \{\mathcal{D} \in \mathcal{M} : \mathcal{D}^* \not\subset \mathcal{D}\}$. Since $\mathcal{M}_1 \cup \mathcal{M}_2 = \mathcal{M}\backslash\{\mathcal{D}^*\}$, it suffices to show

$$\lim_{n\to\infty} \mathrm{pr}\Big\{ \min_{\mathcal{D}\in\mathcal{M}_1} \mathrm{MQBIC}(\mathcal{D}) > \mathrm{MQBIC}(\mathcal{D}^*)\Big\} = 1, \tag{B.17}$$

$$\lim_{n\to\infty} \mathrm{pr}\Big\{ \min_{\mathcal{D}\in\mathcal{M}_2} \mathrm{MQBIC}(\mathcal{D}) > \mathrm{MQBIC}(\mathcal{D}^*)\Big\} = 1. \tag{B.18}$$

We first show (B.17). Write

$$\widehat{W}_{\mathcal{D}} = n^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}}),$$

$$W^* = n^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n} \rho_m(\varepsilon_{kmi}).$$

From Lemma B.2 we know that we can choose some sequence L_n that does not depend on $\mathcal{D}$ and satisfies $L_n \to \infty$, $L_n = o(T_n)$ and $n^{-1}L_n d_n \log n \to 0$, such that for $k = 1, \ldots, K$ and $m = 1, \ldots, M$,

$$\begin{aligned}\mathrm{pr}\Big\{ \Big| \sum_{i=1}^{n} \{\rho_m(Y_i - X_{ki\mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{km\mathcal{D}}) - \rho_m(\varepsilon_{kmi})\} \Big| \\ \leq (MK)^{-1} L_n |\mathcal{D}| \log n, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} \to 1. \end{aligned} \tag{B.19}$$

Then since

$$|\widehat{W}_{\mathcal{D}} - W^*| \leq n^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \Big| \sum_{i=1}^{n} \{\rho_m(Y_i - X_{ki\mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{km\mathcal{D}}) - \rho_m(Y_i - X_{ki\mathcal{D}^*}^{\mathrm{T}} \theta^*_{km\mathcal{D}^*})\} \Big|,$$

we have

$$\mathrm{pr}\Big\{ |\widehat{W}_{\mathcal{D}} - W^*| \leq n^{-1} L_n |\mathcal{D}| \log n, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} \to 1.$$

It follows that

$$\mathrm{pr}\Big\{ |\widehat{W}_{\mathcal{D}} - \widehat{W}_{\mathcal{D}^*}| \leq n^{-1} L_n (|\mathcal{D}| + |\mathcal{D}^*|) \log n, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} \to 1, \tag{B.20}$$

and that, for some postive constants w_1 and w_2,

$$\mathrm{pr}\Big\{ w_1 \leq \widehat{W}_{\mathcal{D}^*} \leq w_2, \text{ for any } \mathcal{D} \in \mathcal{M}_1^* \Big\} \to 1 \tag{B.21}$$

using Assumption 3.9 and the fact that $n^{-1} L_n |\mathcal{D}^*| \log n \to 0$ (Assumptions 3.7). Therefore, with probability tending to one,

$$\begin{aligned} &\min_{\mathcal{D} \in \mathcal{M}_1} \text{MQBIC}(\mathcal{D}) - \text{MQBIC}(\mathcal{D}^*) \\ &= \min_{\mathcal{D} \in \mathcal{M}_1} \Big[\log\{1 + \widehat{W}_{\mathcal{D}^*}^{-1} (\widehat{W}_{\mathcal{D}} - \widehat{W}_{\mathcal{D}^*})\} + (2n)^{-1} T_n (|\mathcal{D}| - |\mathcal{D}^*|) \log n \Big] \end{aligned}$$

$$\geq \min_{\mathcal{D}\in\mathcal{M}_1}\Big\{ -2\widehat{W}_{\mathcal{D}^*}^{-1}|\widehat{W}_{\mathcal{D}}-\widehat{W}_{\mathcal{D}^*}| + (2n)^{-1}T_n(|\mathcal{D}|-|\mathcal{D}^*|)\log n\Big\}$$
$$\geq \min_{\mathcal{D}\in\mathcal{M}_1}\Big\{ -cn^{-1}L_n(|\mathcal{D}|+|\mathcal{D}^*|)\log n + (2n)^{-1}T_n(|\mathcal{D}|-|\mathcal{D}^*|)\log n\Big\}. \quad \text{(B.22)}$$

The first inequality in the above derivation comes from the fact that $\log(1+x) \geq -2|x|$ for any $|x| \in (-1/2, 1/2)$, from (B.20) combined with the fact that $n^{-1}L_n d_n \log n \to 0$, and from (B.21). The last step holds true because of (B.20) and (B.21). Then (B.22) implies (B.17) because $L_n = o(T_n)$ and $|\mathcal{D}| > |\mathcal{D}^*|$.

To prove (B.18), define $\mathcal{D}' = \mathcal{D} \cup \mathcal{D}^*$ for any $\mathcal{D} \in \mathcal{M}_2$. Since q is fixed by Assumption 3.7, we can take

$$\nu = \min_{1\leq k\leq K}\min_{1\leq m\leq M}\min_{j\in\mathcal{D}^*}|\theta^*_{kmj}| > 0,$$

i.e., the smallest value of the nonzero parameters. Since (B.3) still holds for any set in $\mathcal{M}_2^* = \{\mathcal{D} \subset \{1,\ldots,p_n\} : |\mathcal{D}| \leq 2d_n, \mathcal{D}^* \subset \mathcal{D}\}$, we have

$$\mathrm{pr}\Big\{\max_{\mathcal{D}\in\mathcal{M}_2}\|\widehat{\theta}_{km\mathcal{D}'} - \theta^*_{km\mathcal{D}'}\| \leq \nu\Big\} \to 1. \quad \text{(B.23)}$$

For $k = 1,\ldots,K$, $m = 1,\ldots,M$ and any $\mathcal{D} \in \mathcal{M}_2$, let $\widetilde{\theta}_{km\mathcal{D}'}$ be a $|\mathcal{D}'| \times 1$ vector, i.e., the dimension of $\widetilde{\theta}_{km\mathcal{D}'}$ is given by the number of indices in the set $\mathcal{D}' = \mathcal{D} \cup \mathcal{D}^*$. We define it as an extended version of $\widehat{\theta}_{km\mathcal{D}}$: the components of $\widetilde{\theta}_{km\mathcal{D}'}$ that correspond to the index set $\mathcal{D}$ coincide with the components of $\widehat{\theta}_{km\mathcal{D}}$; the remaining components are filled with zeros. For example, if $\mathcal{D} = \{1,3\}$, $\mathcal{D}^* = \{1,2\}$ and $\widehat{\theta}_{km\mathcal{D}} = \{1.4, 0.7\}$, then $\mathcal{D}' = \{1,2,3\}$, $|\mathcal{D}'| = 3$ and $\widetilde{\theta}_{km\mathcal{D}'} = (1.4, 0, 0.7)^{\mathrm{T}}$. Since $\mathcal{D}^* \not\subset \mathcal{D}$, there exist some k_0 and m_0 such that $\|\widetilde{\theta}_{k_0m_0\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'}\| \geq \nu$. Combined with (B.23) and since the check function is convex, this implies that there exists a $|\mathcal{D}'| \times 1$ vector $\bar{\theta}_{\mathcal{D}'}$ such that $\|\bar{\theta}_{\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'}\| = \nu$ and

$$\sum_{i=1}^n \rho_{m_0}(Y_{k_0i} - X^{\mathrm{T}}_{k_0i\mathcal{D}'}\bar{\theta}_{\mathcal{D}'}) \leq \sum_{i=1}^n \rho_{m_0}(Y_{k_0i} - X^{\mathrm{T}}_{k_0i\mathcal{D}'}\widetilde{\theta}_{k_0m_0\mathcal{D}'}) = \sum_{i=1}^n \rho_{m_0}(Y_{k_0i} - X^{\mathrm{T}}_{k_0i\mathcal{D}}\widehat{\theta}_{k_0m_0\mathcal{D}}).$$

Write

$$B_\nu(\mathcal{D}') = \{\omega \in \mathbb{R}^{|\mathcal{D}'|} : \|\omega\| = \nu\},$$
$$G_{\mathcal{D}'}(\omega) = n^{-1}\sum_{i=1}^{n}\{\rho_{m_0}(\varepsilon_{k_0m_0i} - X_{k_0i\mathcal{D}'}^{\mathrm{T}}\omega) - \rho_{m_0}(\varepsilon_{k_0m_0i})\}.$$

Then we have, for any $\mathcal{D} \in \mathcal{M}_2$,

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\{\rho_{m_0}(Y_{k_0i} - X_{k_0i\mathcal{D}}^{\mathrm{T}}\widehat{\theta}_{k_0m_0\mathcal{D}}) - \rho_{m_0}(Y_{k_0i} - X_{k_0i\mathcal{D}'}^{\mathrm{T}}\widehat{\theta}_{k_0m_0\mathcal{D}'})\} \\
\geq\ & n^{-1}\sum_{i=1}^{n}\{\rho_{m_0}(Y_{k_0i} - X_{k_0i\mathcal{D}}^{\mathrm{T}}\bar{\theta}_{\mathcal{D}'}) - \rho_{m_0}(Y_{k_0i} - X_{k_0i\mathcal{D}'}^{\mathrm{T}}\widehat{\theta}_{k_0m_0\mathcal{D}'})\} \\
=\ & G_{\mathcal{D}'}(\bar{\theta}_{\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'}) - G_{\mathcal{D}'}(\widehat{\theta}_{k_0m_0\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'}) + \\
& E\{G_{\mathcal{D}'}(\bar{\theta}_{\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'})|X_{k_0\cdot\mathcal{D}'}\} - E\{G_{\mathcal{D}'}(\bar{\theta}_{\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'})|X_{k_0\cdot\mathcal{D}'}\} \\
\geq\ & \inf_{\omega\in B_\nu(\mathcal{D}')} E\{G_{\mathcal{D}'}(\omega)|X_{k_0\cdot\mathcal{D}}\} - \sup_{\omega\in B_\nu(\mathcal{D}')}\Big|G_{\mathcal{D}'}(\omega) - E\{G_{\mathcal{D}'}(\omega)|X_{k_0\cdot\mathcal{D}'}\}\Big| - \\
& G_{\mathcal{D}'}(\widehat{\theta}_{k_0m_0\mathcal{D}'} - \theta^*_{k_0m_0\mathcal{D}'}).
\end{aligned} \tag{B.24}$$

Similar to the calculation of (B.7), we have for any $\mathcal{D}' \in \mathcal{M}_2^*$ and $\omega \in B_\nu(\mathcal{D}')$,

$$\begin{aligned}
E\{G_{\mathcal{D}'}(\omega)|X_{k_0\cdot\mathcal{D}'}\} &= n^{-1}\sum_{i=1}^{n}\int_0^{X_{k_0i\mathcal{D}'}^{\mathrm{T}}\omega} F_{k_0m_0}(s|X_{k_0i\mathcal{D}'}) - F_{k_0m_0}(0|X_{k_0i\mathcal{D}'})\,ds \\
&= n^{-1}\sum_{i=1}^{n}\int_0^{X_{k_0i\mathcal{D}'}^{\mathrm{T}}\omega} s f_{k_0m_0}(\bar{s}|X_{k_0i\mathcal{D}'})\,ds \\
&\geq c\,\omega^{\mathrm{T}}\Big\{n^{-1}\sum_{i=1}^{n}(X_{k_0i\mathcal{D}'}X_{k_0i\mathcal{D}'}^{\mathrm{T}})\Big\}\omega \\
&\geq c\lambda_{\min}(n^{-1}X_{k_0\cdot\mathcal{D}'}^{\mathrm{T}}X_{k_0\cdot\mathcal{D}'})\|\omega\|^2 = c\|\omega\|^2
\end{aligned} \tag{B.25}$$

where the third step uses Assumption (3.3) and the last step uses Assumption (3.6). Then under Assumptions 3.1, 3.3, 3.6 and 3.7, Lemma A.3 of the supplement to Lee et al. (2014) gives

$$\max_{\mathcal{D}'\in\mathcal{M}_2^*}\sup_{\omega\in B_\nu(\mathcal{D}')}\Big|G_{\mathcal{D}'}(\omega) - E\{G_{\mathcal{D}'}(\omega)|X_{k_0\cdot\mathcal{D}'}\}\Big| = o_p(1). \tag{B.26}$$

It is obvious that (B.19) is still valid when $\mathcal{M}_1^*$ is subsitutued by $\mathcal{M}_2^*$. Hence

$$\mathrm{pr}\Big\{ \max_{\mathcal{D}' \in \mathcal{M}_2^*} |G_{\mathcal{D}'}(\widehat{\theta}_{k_0 m_0 \mathcal{D}'} - \theta^*_{k_0 m_0 \mathcal{D}'})| \le c n^{-1} L_n d_n \log n \Big\} \to 1,$$

which gives

$$\max_{\mathcal{D}' \in \mathcal{M}_2^*} \Big| G_{\mathcal{D}'}(\widehat{\theta}_{k_0 m_0 \mathcal{D}'} - \theta^*_{k_0 m_0 \mathcal{D}'}) \Big| = o_p(1).$$

This, combined with (B.24), (B.25) and (B.26) implies that, with probability approaching one,

$$n^{-1} \min_{\mathcal{D} \in \mathcal{M}_2} \sum_{i=1}^{n} \{ \rho_m(Y_{k_0 i} - X_{k_0 i \mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{k_0 m_0 \mathcal{D}}) - \rho_m(Y_{k_0 i} - X_{k_0 i \mathcal{D}'} \widehat{\theta}_{k_0 m_0 \mathcal{D}'}) \} \ge 2C. \tag{B.27}$$

Since $\mathcal{D} \in \mathcal{D}'$, we have

$$\sum_{i=1}^{n} \{ \rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{km\mathcal{D}}) - \rho_m(Y_{ki} - X_{ki\mathcal{D}'} \widehat{\theta}_{km\mathcal{D}'}) \} \ge 0$$

for any k, m and $\mathcal{D} \in \mathcal{M}_2$. It follows

$$\begin{aligned} \widehat{W}_{\mathcal{D}} - \widehat{W}_{\mathcal{D}'} &= n^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i=1}^{n} \{ \rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{km\mathcal{D}}) - \rho_m(Y_{ki} - X_{ki\mathcal{D}'} \widehat{\theta}_{km\mathcal{D}'}) \} \\ &\ge n^{-1} \sum_{i=1}^{n} \{ \rho_m(Y_{k_0 i} - X_{k_0 i \mathcal{D}}^{\mathrm{T}} \widehat{\theta}_{k_0 m_0 \mathcal{D}}) - \rho_m(Y_{k_0 i} - X_{k_0 i \mathcal{D}'} \widehat{\theta}_{k_0 m_0 \mathcal{D}'}) \}. \end{aligned}$$

This, combined with (B.27), gives

$$\mathrm{pr}\Big\{ \min_{\mathcal{D} \in \mathcal{M}_2} (\widehat{W}_{\mathcal{D}} - \widehat{W}_{\mathcal{D}'}) \ge 2C \Big\} \to 1 \tag{B.28}$$

Then with probability tending to one,

$$\min_{\mathcal{D} \in \mathcal{M}_2} \mathrm{MQBIC}(\mathcal{D}) - \mathrm{MQBIC}(\mathcal{D}')$$

$$
\begin{aligned}
&= \min_{\mathcal{D}\in\mathcal{M}_2} \left[\log\{1 + \widehat{W}_{\mathcal{D}'}^{-1}(\widehat{W}_{\mathcal{D}} - \widehat{W}_{\mathcal{D}'})\} - (2n)^{-1} T_n(|\mathcal{D}'| - |\mathcal{D}|) \log n \right] \\
&\geq \min_{\mathcal{D}\in\mathcal{M}_2} \left[\min\{\log 2, \widehat{W}_{\mathcal{D}'}^{-1}(\widehat{W}_{\mathcal{D}} - \widehat{W}_{\mathcal{D}'})/2\} - (2n)^{-1} T_n |\mathcal{D}^*| \log n \right] \\
&\geq \min_{\mathcal{D}\in\mathcal{M}_2} \left[\min\{\log 2, \widehat{W}_{\mathcal{D}'}^{-1} C\} - (2n)^{-1} T_n |\mathcal{D}^*| \log n \right] > 0 \qquad \text{(B.29)}
\end{aligned}
$$

The first inequality comes from the fact that $\log(1 + x) \geq \min\{x/2, \log 2\}$ for any $x \geq 0$. The second inequality uses (B.28). The last step uses Assumption 3.8 and the fact that (B.21) is still valid when $\mathcal{M}_1^*$ is substituted by $\mathcal{M}_2^*$. Since (B.17) can be easily extended to any $\mathcal{D} \in (\mathcal{M}_2^* \backslash \{\mathcal{D}^*\})$, we know that, with probability tending to one, MQBIC($\mathcal{D}'$) $\geq$MQBIC($\mathcal{D}^*$) for any $\mathcal{D}' \in \mathcal{M}_2^*$. This and (B.29) yield

$$
\begin{aligned}
&\min_{\mathcal{D}\in\mathcal{M}_2} \text{MQBIC}(\mathcal{D}) - \text{MQBIC}(\mathcal{D}^*) \\
&= \min_{\mathcal{D}\in\mathcal{M}_2} \{\text{MQBIC}(\mathcal{D}) - \text{MQBIC}(\mathcal{D}') + \text{MQBIC}(\mathcal{D}') - \text{MQBIC}(\mathcal{D}^*)\} \\
&\geq \min_{\mathcal{D}\in\mathcal{M}_2} \{\text{MQBIC}(\mathcal{D}) - \text{MQBIC}(\mathcal{D}')\} > 0
\end{aligned}
$$

with probability tending to one. This proves (B.18).